ENCYCLOPEDIA OF

Physical Science and Technology

THIRD EDITION

Biotechnology

AP

Table of Contents
(Subject Area: Biotechnology)

Article	Authors	Pages in the Encyclopedia
Biomass Utilization, Limits of	*David Pimentel*	Pages 159-171
Biomass, Bioengineering of	*Bruce E. Dale*	Pages 141-157
Biomaterials, Synthetic Synthesis, Fabrication, and Applications	*Carole C. Perry*	Pages 173-191
Biomineralization and Biomimetic Materials	*Paul D. Calvert*	Pages 193-205
Bioreactors	*Yusuf Chisti and Murray Moo-Young*	Pages 247-271
Fiber-Optic Chemical Sensors	*David R. Walt, Israel Biran and Tarun K. Mandal*	Pages 803-829
Hybridomas, Genetic Engineering of	*Michael Butler*	Pages 427-443
Image-Guided Surgery	*Ferenc A. Jolesz*	Pages 583-594
Mammalian Cell Culture	*Bryan Griffiths and Florian Wurm*	Pages 31-47
Metabolic Engineering	*Jens Nielsen*	Pages 391-406
Microanalytical Assays	*Jerome S. Schultz*	Pages 679-694
Optical Fiber Techniques for Medical Applications	*Abraham Katzir*	Pages 315-333
Pharmaceuticals, Controlled Release of	*Giancarlo Santus and Richard W. Baker*	Pages 791-803
Pharmacokinetics	*Michael F. Flessner*	Pages 805-820
Separation and Purification of Biochemicals	*Laure G. Berruex and Ruth Freitag*	Pages 651-673
Tissue Engineering	*François Berthiaume and Martin L. Yarmush*	Pages 817-842
Toxicology in Forensic Science	*Olaf H. Drummer*	Pages 905-911

Biomass Utilization, Limits of

David Pimentel
Cornell University

I. Biomass Resources
II. Conversion of Biomass Resources
III. Biogas
IV. Biomass and the Environment
V. Social and Economic Impacts
VI. Conclusion

GLOSSARY

Biodiversity All species of plants, animals, and microbes in one ecosystem or world.

Biogas A mixture of methane and carbon dioxide produced by the bacterial decomposition of organic wastes and used as a fuel.

Biomass Amount of living matter, including plants, animals, and microbes.

Energy Energy is the capacity to do work and includes heat, light, chemical, acoustical, mechanical, and electrical.

Erosion The slow breakdown of rock or the movement and transport of soil from one location to another. Soil erosion in crop and livestock production is considered serious worldwide.

Ethanol Also called ethyl alcohol. A colorless volatile flammable liquid with the chemical formula C_2H_5OH that is the intoxicating agent in liquors and is also used as a solvent.

Methanol Also called methyl alcohol. A light volatile flammable liquid with the chemical formula CH_3OH that is used especially as a solvent, antifreeze, or denaturant for ethyl alcohol and in the synthesis of other chemicals.

Pollution The introduction of foreign, usually man-made, products or waste into the environment.

Pyrolysis Chemical change brought about by the action of heat.

Subsidy A grant or gift of money.

THE INTERDEPENDENCY of plants, animals, and microbes in natural ecosystems has survived well for billions of years even though they only captured 0.1% of the sun's energy. All the solar energy captured by vegetation and converted into plant biomass provides basic resources for all life, including humans. Approximately 50% of the world's biomass is used by humans for food plus lumber and pulp and medicines, as well as support for all other animals and microbes in the natural ecosystem. In addition some biomass is converted into fuel.

Serious shortages of biomass for human use and maintaining the biodiversity in natural ecosystems now exist throughout the world. Consider that more than 3 billion humans are now malnourished, short of food, and various

essential nutrients. This is the largest number and proportion of malnourished humans ever recorded in history. Meanwhile, based on current rates of increase, the world population is projected to double to more than 12 billion in approximately 50 years. With a population growth of this magnitude, the numbers of malnourished could reach 5 billion within a few decades. The need for biomass will continue to escalate.

Associated with increasing human numbers are diverse environmental problems, including deforestation, urbanization, industrialization, and chemical pollution. All these changes negatively impact on biomass production that is vital to human life and biodiversity. However, at present and in the foreseeable future the needs of the rapidly growing human population will stress biomass supplies. In our need to supply food and forest products for humans from biomass, intense competition between human needs for food and the conversion of biomass into an energy resource is expected to intensify in the coming decades.

Furthermore, human intrusion throughout the natural environment is causing a serious loss of biodiversity with as many as 150 species being lost per day. The present rate of extinction of some groups of organisms is 1000–10,000 times faster than that in natural systems. Ecosystem and species diversity are the vital reservoir of genetic material for the successful development of agriculture, forestry, pharmaceutical products, and biosphere services in the future.

The limits of biomass energy utilization and how this relates to food production and natural biodiversity and environmental quality are discussed in this article.

I. BIOMASS RESOURCES

The amount of biomass available is limited because plants on average capture only about 0.1% of the solar energy reaching the earth. Temperature, water availability, soil nutrients, and feeding pressure of herbivores all limit biomass production in any given region. Under optimal growing conditions, natural and agricultural vegetation and produce about 12 million kilocalories per hectare per year (about 3 t/ha dry biomass).

A. World Biomass

The productive ecosystems in the world total an estimated 50 billion hectare, excluding the icecaps. Marine ecosystems occupy approximately 36.5 billion hectare while the terrestrial ecosystems occupy approximately 13.5 billion hectare. Gross primary productivity for the marine ecosystem is estimated to be about 1 t/ha/yr, making the total biomass production about 36.5 billion metric tons or 145×10^{15} kcal/yr. In contrast, the terrestrial ecosystem produces about 3 t/ha/yr, making the total biomass about 40.5 billion tons or 162×10^{15} kcal/yr. The total biomass produced is approximately 77 billion tons or about 12.8 t per person per year.

The 40.5 billion tons of biomass produced in the terrestrial ecosystem provides an estimated 6.8 t/yr per person. Given that humans harvest about 50% of the world's terrestrial biomass, each person is utilizing 3.4 t/yr. This 3.4 t/yr includes all of agriculture, including livestock production and forestry. The remaining 3.4 t/yr per person supplies the other 10 million species of natural biota their energy and nutrient needs.

Currently, approximately 50% of the world's biomass (approximately 600 quads worldwide) is being used by humans for food, construction, and fuel. This major utilization of biomass, habitat destruction associated with the rapid increase in the world population, and environmental pollution from about 100,000 chemicals used by humans is causing the serious loss of biodiversity worldwide. With each passing day an estimated 150 species are being eliminated because of increasing human numbers and associated human activities, including deforestation, soil and water pollution, pesticide use, urbanization, and industrialization.

B. United States Biomass

In the North American temperate region, the solar energy reaching a hectare of land per year is 14 billion kilocalories. However, plants do not grow during the winter there. Most plant growth occurs during 4 months in the summer when about 7 billion kilocalories reach a hectare. In addition to low temperatures, plant growth is limited by shortages of water, nitrogen, phosphorus, potassium, and other nutrients, plus the feeding pressure of herbivores and disease organisms. At most, during a warm moist day in July a plant, like corn, under very favorable conditions, might capture only 5% of the sunlight energy reaching the plants. Under natural and agricultural conditions for the total year, vegetation produces approximately 12 million kilocalories per hectare per year or about 3 t/ha dry biomass.

Total annual biomass produced in the United States is an estimated 2.6 billion tons (Table I). This is slightly more than 6% of all the terrestrial biomass produced in the world. Based on the United States. land area of 917 million hectares, this is the equivalent of 2.9 t/ha/yr and is similar to the world average of 3 t/ha/yr for all the terrestrial ecosystems of the world. The total energy captured by all the United States plant biomass each year is approximately 11.8×10^{15} kcal (Table I). With the United States currently consuming 87 quads (21.8×10^{15} kcal)

TABLE I Annual Biomass Production in the United States

	Land area (10^6/ha)	Biomass production (10^6/t)
Cropland and crops	192	1,083
Pasture and forage	300	900
Forests	290	580
Other	135	68
Total area	917	—
Total biomass	—	2,631
Total energy (10^{15}/kcal)	11.8	
Biomass production (t/ha)	2.9	

[From Pimentel, D., and Kounang, N. (1998), *Ecosystems* **1,** 416–426.]

of fossil energy each year, this means that it is consuming 85% more fossil energy than the total energy captured by all its plant biomass each year.

C. United States Agricultural and Forest Products and Biofuels

Including crops and forages from pastures, the United States harvests approximately 1307 million tons of biomass per year in agricultural products and approximately 100 million tons of biomass per year as forest products (Table II). Together the energy value of harvested agricultural and forest products total 6352 10^{12} kcal/yr (Table II). These data suggest that the United States is harvesting in the form of agricultural and forest products, 54% of the total energy captured each year by the United States biomass annually (Tables I and II). This total does not include the biomass harvested now and used as biofuel.

II. CONVERSION OF BIOMASS RESOURCES

In addition to using biomass directly as food, fiber, lumber, and pulp, biomass is utilized as a fuel. The total biofuel utilized in the United States is slightly more than 3 quads (800×10^{12} kcal) per year. If the biofuel energy is added to that harvested as agricultural and forest products, then the total biomass energy harvested from the United States terrestrial ecosystem is 7332×10^{12} kcal/yr. This is equivalent to 62% of the total biomass energy produced in the United States each year. Harvesting this 62% is having a negative impact on biodiversity in the nation.

A. Direct Heating

Heat production is the most common conversion system for using biomass resources. Heat from wood and other biomass resources is utilized for cooking food, heating homes, and producing steam for industry.

Each year, worldwide, an estimated 5300 million dry tons of biomass are burned directly as a fuel, providing about 88 quads of energy. Rural poor in developing countries obtain up to 90% of their energy needs by burning biomass. In developing countries, about 2 billion tons of fuelwood, 1.3 billion tons of crop residues, plus nearly 1 billion tons of dung are burned each year.

Although some deforestation results from the use of fuelwood, the most significant environmental impacts result from burning crop residues and dung. When crop residues and dung are removed from the land and used as a fuel this leaves the cropland without vegetative protection and exposed to wind and water erosion. Erosion destroys the productivity of cropland, by robbing the soil of nutrients, essential water, soil organic matter, and adequate rooting depth.

Cooking requires relatively large amounts of fuel and is essential for preventing disease, improving nutrition, and increasing the palatability of many foods. The transfer of heat from the woodfire in a stove to the food product is about 33% efficient, while over an open fire, the heat transfer to the food is only about 10% efficient. Under usual cooking conditions, from 2 to 3 kcal are required to cook 1 kcal of food.

TABLE II Total Annual Amount of Solar Energy Harvested in the Form of Agricultural and Forest Biomass in the U.S.

	Tons (10^6)	Energy (10^{12} kcal)
Corn	194	873
Wheat	71	320
Rice	6	27
Soybeans	51	230
Sorghum	22	99
Potatoes	16	72
Vegetables	6	27
Fruits	5	23
Nuts	0.8	4
Oil seeds	9	41
Sugarcane	2.5	20
Sugar beets	2	27
Pulses	1	5
Oats	7	32
Rye	1	5
Barley	13	59
Total	**407.3**	**1,853**
Pasture forage	900	4,050
Forest products	100	450
Totals	**1,407**	**6,352**
Total per capita (tons)	5.2	
Total per capita (10^6/kcal)	23.3	

[From Pimentel, D., and Kounang, N. (1998), *Ecosystems* **1,** 416–426.]

In a developing country an average, 600–700 kg/yr of dry biomass per person is used for cooking. For example, the use of fuelwood for cooking and heating in Nepal is about 846 kg/yr of biomass per person. Other investigators report that from 912 to 1200 kg/yr of biomass per person is used for both cooking and heating. In some developing countries, fuelwood for cooking and heating may cost almost as much as the food, making it necessary to use crop residues and dung.

A significant amount of wood is converted into charcoal for cooking and heating. Similar to wood fires for cooking, open charcoal fires are only about 10% efficient in transferring heat energy to food. However, charcoal has some advantages over wood. First, it is lightweight and easy to transport. One kilogram of charcoal contains about 7100 kcal of potential energy in contrast to a kilogram of wood that has about 4000 kcal. Charcoal burns more uniformly and with less smoke than wood.

However, charcoal production is an energy-intensive process. Although charcoal has a high energy content, from 20,300 to 28,400 kcal of hardwood must be processed to obtain the 7100 kcal of charcoal. Considering this low conversion efficiency ranging from 25 to 35%, charcoal heating for cooking has an overall energy transfer efficiency to food of only 2.5–3.5%. Further, the use of charcoal uses more forest biomass than directly burning the wood.

Using fuelwood for the production of steam in a boiler under relatively optimal conditions is 55–60% efficient, that is, burning 4000 kcal of air-dried wood provides from 2200 to 2400 kcal of steam in the boiler. More often the efficiency is less than 55–60%. Steam production is used to produce electricity and producing a salable product, such as steam, for industrial use.

Collecting biomass for fuel requires a substantial amount of time and human effort. For example, in Indonesia, India, Ghana, Mozambique, and Peru families spend from 1.5 to 5 hrs each day collecting biomass to use as a fuel.

Estimates are that more than half of the people who depend on fuelwood have inadequate supplies. In some countries, such as Brazil, where forest areas are at present fairly abundant, the rural poor burn mostly wood and charcoal. However, in many developing countries crop residues account for most of the biomass fuel, e.g., 55% in China, 77% in Egypt, and 90% in Bangladesh. Estimates are that the poor in these countries spend 15–25% of their income for biomass fuel.

B. Health Effects

Environmentally, burning biomass is more polluting than using natural gas, but less polluting than coal. Biomass combustion releases more than 200 different chemical pollutants into the atmosphere. The pollutants include, up to 14 carcinogens, 4 cocarcinogens, and 6 toxins that damage cilia, plus additional mucus-coagulating agents. Wood smoke contains pollutants known to cause bronchitis, emphysema, cancer, and other serious illnesses.

Globally, but especially in developing nations where people cook with fuelwood over open fires, approximately 4 billion humans suffer continuous exposure to smoke. This smoke which contains large quantities of particulate matter and more than 200 chemicals, including several carcinogens, results in pollution levels that are considerably above those acceptable by the World Health Organization (WHO). Worldwide fuelwood smoke is estimated to cause the death of 4 million children each year worldwide. In India, where people cook with fuelwood and dung, particulate concentrations in houses are reported to range from 8300 to 15,000 $\mu g/m^3$, greatly exceeding the 75 $\mu g/m^3$ maximum standard for indoor particulate matter in the United States.

Because of the release of pollutants, some communities in developed areas, such as Aspen, CO, have banned wood burning for heating homes. When biomass is burned continuously in a confined space for heating, its pollutants accumulate and can become a serious health threat.

C. Ethanol Production

Numerous studies have concluded that ethanol production does not enhance energy security, is not a renewable energy source, is not an economical fuel, and does not insure clean air. Further, its production uses land suitable for crop production and causes environmental degradation.

The conversion of corn and other food/feed crops into ethanol by fermentation is a well-known and established technology. The ethanol yield from a large plant is about 9.5 l (2.5 gal) from a bushel of corn of 24.5 kg (2.6 kg/l of ethanol). Thus, a hectare of corn yielding 7965 kg/ha could be converted into about 3063 l of ethanol.

The production of corn in the United States requires a significant energy and dollar investment (Table III). For example, to produce 7965 kg/ha of corn using conventional production technology requires the expenditure of about 10.4 million kcal (about 10,000 l of oil equivalents) (Table III), costing about $857.17 for the 7965 kg or approximately 10.8¢/kg of corn produced. Thus, for a liter of ethanol, the corn feedstock alone costs 28¢.

The fossil energy input to produce the 7965 kg/ha corn feedstock is 10.4 million kilocalories or 3408 kcal/l of ethanol (Table III). Although only 16% of United States corn production is currently irrigated, it is included in the analysis, because irrigated corn production is energy costly. For the 150 mm of irrigation water applied and

TABLE III Energy Inputs and Costs of Corn Production per Hectare in the United States

Inputs	Quantity	kcal × 1000	Costs
Labor	11.4 hr	561	$100.00
Machinery	55 kg	1,018	103.21
Diesel	42.2 L	481	8.87
Gasoline	32.4 L	328	9.40
Nitrogen	144.6 kg	2,668	89.65
Phosphorus	62.8 kg	260	34.54
Potassium	54.9 kg	179	17.02
Lime	699 kg	220	139.80
Seeds	21 kg	520	74.81
Herbicides	3.2 kg	320	64.00
Insecticides	0.92 kg	92	18.40
Irrigation	150 mm	3,072	150.00
Electricity	13.2 kg	34	2.38
Transportation	151 kg	125	45.30
Total		10,439	$857.17
Corn yield		27,758	
= 7,965 kg		kcal output/kcal input = 1 : 2.66	

From Pimentel, D., Doughty, R., Carothers, C., Lamberson, S., Bora, N., and Lee, K. *J. Agr. Environ. Ethics* (in press).

pumped from only 30.5 m (100 feet), the average energy input is 3.1 million kilocalories/hectare (Table III).

When investigators ignore some of the energy inputs in biomass production and processing they reach an incomplete and deficient analysis for ethanol production. In a recent USDA report, no energy inputs were listed for machinery, irrigation, or for transportation. All of these are major energy input costs in United States corn production (Table III). Another way of reducing the energy inputs for ethanol production is to arbitrarily select lower production costs for the inputs. For instance, Shapouri et al. list the cost of a kilogram of nitrogen production at 12,000 kcal/kg, considerably lower than Food and Agricultural Organization of the UN (FAO), which list the cost of nitrogen production at 18,590 kcal/kg. Using the lower figure reduces the energy inputs in corn production by about 50%. Other workers have used a similar approach to that of Shapouri et al.

The average costs in terms of energy and dollars for a large (240 to 280 million liters per year), modern ethanol plant are listed in Table IV. Note the largest energy inputs are for corn production and for the fuel energy used in the fermentation/distillation process. The total energy input to produce 1000 l of ethanol is 8.7 million kilocalories (Table IV). However, 1000 l of ethanol has an energy value of only 5.1 million kilocalories. Thus, there is a net energy loss of 3.6 million kilocalories per 1000 l of ethanol produced. Put another way, about 70% more energy is required to produce 1000 l of ethanol than the energy that actually is in the ethanol (Table IV).

In the distillation process, large amounts of fossil energy are required to remove the 8% ethanol out of the 92% water. For example, to obtain 1000 l of pure ethanol with an 8% ethanol concentration out of 92% water, then this ethanol must come from the 12,500 l of ethanol/water mixture. A total of 124 l of water must be eliminated per liter of ethanol produced. Although ethanol boils at about 78°C, in contrast to water at 100°C, the ethanol is not extracted from the water in one distillation process. Instead, about 3 distillations are required to obtain the 95% pure ethanol that can be mixed with gasoline. To be mixed with gasoline, the 95% ethanol must be further processed with more energy inputs to achieve 99.8% pure ethanol. The three distillations account for the large quantities of fossil energy that are required in the fermentation/distillation process. Note, in this analysis all the added energy inputs for fermentation/distillation process are included, not just the fuel for the distillation process itself.

This contrasts with Shapouri et al. who, in 1995, give only one figure for the fermentation/distillation process and do not state what the 3.4 million kilocalories represents in their analysis for producing 1000 l of ethanol. Careful and detailed analyses and full accountings are needed to ascertain the practicality of ethanol production as a viable energy alternative.

About 61% of the cost of producing ethanol (46¢ per liter) in such a large-production plant is for the corn substrate itself (28¢/l) (Table IV). The next largest input is for coal to fuel the fermentation/distillation process, but this was only 4¢ (Table IV). These ethanol production costs include a small charge for pollution control (6¢ per liter), which is probably a low estimate. In smaller plants with an annual production of 150,000 l/yr, the cost per liter increases to as much as 66¢ per liter. Overall, the per liter

TABLE IV Inputs per 1000 l of Ethanol Produced from Corn

Inputs	Kilograms	Kilocalories (1000)	Dollars
Corn	2,600	3,408	$280
Transport of corn	2,600	312	32
Water	160,000	90	20
Stainless steel	6	89	10
Steel	12	139	10
Cement	32	60	10
Coal	660	4,617	40
Pollution control costs	—	—	60
Total		8,715	$462

From Pimentel, D., Warneke, A. F., Teel, W. S., Schwab, K. A., Simcox, N. J., Ebert, D. M., Baenisch, K. D., and Aaron, M. R., (1988). *Adv. Food. Res.* **32,** 185–238.

price for ethanol does not compare favorably with that for the production of gasoline fuels which presently is about 25¢ per liter.

Based on current ethanol production technology and recent oil prices, ethanol still costs substantially more to produce in dollars than it is worth on the market. Clearly, without the approximately $1 billion subsidy, United States ethanol production would be reduced or cease, confirming the fact that basically ethanol production is uneconomical. Federal subsidies average 16¢ per liter and state subsidies average 5¢ per liter. Because of the relatively low energy content of ethanol, 1.5 l of ethanol is the energy equivalent of 1 l of gasoline. This means that the cost of subsidized ethanol is 68¢ per liter. The current cost of producing gasoline is about 25¢ per liter.

At present, federal and state subsidies for ethanol production total about $1 billion per year and are mainly paid to large corporations (calculated from the above data). The costs to the consumer are greater than the $1 billion per year used to subsidize ethanol production because of increased corn prices. The resulting higher corn prices translate into higher meat, milk, and egg prices because currently about 70% of the corn grain is fed to United States livestock. Doubling ethanol production can be expected to inflate corn prices perhaps as much as 1%. Therefore, in addition to paying tax dollars for ethanol subsidies, consumers would be paying significantly higher food prices in the market place. It should be noted that the USDA is proposing to increase the subsidies to the large corporations by about $400 million per year.

Currently about 3.8 billion liters of ethanol are being produced in the United States each year. This amount of ethanol provides only about 1% of the fuel utilized by United States automobiles. To produce the 3.8 billion liters of ethanol we must use about 1.3 million hectares of land. If we produced 10% of United States fuel the land requirement would be 13 million hectares. Moreover not all the 3.8 billion liters would be available to use, because a lot would be needed to sow, fertilize, and harvest 13 million hectares. Clearly, corn is not a renewable resource for ethanol energy production.

The energy and dollar costs of producing ethanol can be offset in part by the by-products produced, especially the dry distillers grains (DDG) made from dry-milling that can be fed primarily to cattle. Wet-milling ethanol plants produce such by-products as corn gluten meal, gluten feed, and oil. Sales of the by-products help offset the energy and economic costs of ethanol production. For example, use of by-products can offset the ethanol production costs by 8–24% (Table IV). The resulting energy output/input comparison, however, remains negative (Table IV). The sales of the by-products that range from13 to 16¢ per liter do not make ethanol competitive with gasoline.

Furthermore, some of the economic and energy contributions of the by-products are negated by the environmental pollution costs associated with ethanol production. These are estimated to be about 6¢ per liter (Table IV). In United States corn production, soil erodes about 12 times faster than it can be reformed. In irrigated corn acreage, ground water is being mined 25% faster than its natural recharge rate. This suggests that the environmental system in which corn is being produced is being rapidly degraded. Further, it substantiates the finding that the United States corn production system is not sustainable for the future, unless major changes are made in the cultivation of this major food/feed crop. Corn should not be considered a renewable resource for ethanol energy production.

When considering the advisability of producing ethanol for automobiles, the amount of cropland required to grow corn to fuel each automobile should be understood. To clarify this, the amount of cropland needed to fuel one automobile with ethanol was calculated. An average United States automobile travels about 16,000 km/yr and uses about 1900 l/yr of gasoline. Although 8000 kg/ha of corn will yield about 3100 l of ethanol, it has an energy equivalent of only 1952 l because ethanol has a much lower kilocalories content than gasoline.

However, even assuming *zero* or no energy charge for the fermentation and distillation process and charging *only* for the energy required to produce corn (Table III), the net fuel energy yield from 1 ha of corn is 433 l. Thus, to provide 1900 l per car, about 4.4 ha of corn must be grown to fuel one car with ethanol for one year. In comparison, only 0.6 ha of cropland is currently used to feed each American. Therefore, more than seven times more cropland would be required to fuel one automobile than is required to feed one American.

Assuming a net production of 433 l of fuel per corn hectare and if all automobiles in the United States were fueled with ethanol, then a total of approximately 900 million hectares of cropland land would be required to provide the corn feedstock for production. This amount of cropland would equal nearly the total land area of the United States.

Brazil had been a large producer of ethanol, but has abandoned subsidizing it. Without the subsidy, economic ethanol production is impossible.

III. BIOGAS

Biomass material that contains large quantities of water can be effectively converted into usable energy using naturally occurring microbes in an anaerobic digestion system. These systems use feedstocks, like dung and certain plants such as water hyacinth, although production and

harvesting costs of the latter are generally greater than for dung. The processing facility can be relatively simple and be constructed for about $700. A large facility capable of processing the dung from 320 cows might cost about $150,000. The basic principles for both systems are similar.

Manure from a dairy farm or small cattle operation is loaded or pumped into a sealed, corrosion-resistant digestion tank where it is held from 14 to 28 days at temperatures from 30 to 38°C. In some digestion systems, the manure in the tank is constantly stirred to speed the digestion process and assure even heating. During this period, the mesophilic bacteria break down volatile solids (VS) in the manure and convert them into methane gas (65%) and carbon dioxide (35%). Small amounts of hydrogen sulfide may also be produced. This gas is drawn off through pipes and either burned directly, similar to natural gas, or scrubbed to clean away the hydrogen sulfide and used to generate electricity. The energy output/input is listed in Table V.

The amount of biogas produced in this system is determined by the temperature of the system, the VS content of the feedstock, and the efficiency of converting it into biogas. This efficiency varies from 18 to 95%. Dairy cows produce 85 kg daily of manure for each 1000 kg of live weight. The total solids in this manure average 10.6 kg, and of these, 8.6 kg are VS. Theoretically, a 100% efficient digester could produce 625 l of biogas for every kilogram of VS in the system. The digester utilized for the data presented in Table V was 28.3% efficient. It produces 177 l of biogas per kilogram of VS added or 1520 l of biogas per 1000 kg live weight of cattle daily. Note, if the total heat value of the manure was used in calculating efficiency, then the percentage efficiency would be only 5%.

Biogas has an energy content of about 5720 kcal/m^3, compared to 8380 kcal/m^3 for pure methane gas, because carbon dioxide is present in the biogas. Energy costs and energy outputs for processing 100 t of manure (wet), with a 7.1 million kilocalories energy input, results in a total of 10.2 million kilocalories produced for a net energy yield of 3.1 million kilocalories (Table V). Much of the energy input or cost comes from the production of electricity to run the pumps and stirring system used to reduce the retention time in the digester. The volume of the digester is determined by the amount of manure produced by the animals during the retention time. In this example, with a retention time of 14 days, it would be slightly over 75 m^3. It is assumed that the electricity is generated from the biogas and that the electrical conversion efficiency of the entire operation is 33%. The energy needed to heat the digester is cogenerated by the electric generator via the use of the generator's cooling system as the heat source. The net energy produced by the digester can either be used to generate electricity for the farm or be used as heat source for other on-farm activities.

Although material costs are lowered if there is no generator or stirring mechanism on the digester, the size of the digester must be increased because of the increased retention time needed to complete the process. Also, some of the biogas will have to be used to heat the digester, perhaps an additional 610,000 kcal for every 100 wet tons of manure digested. The critical heat requirements are calculated by including the heat losses to the surroundings, the heat associated with the feed and effluents, and the heat released by the biological reaction. In the tropics, the overall efficiency of the biogas systems is enhanced because there is no need to heat the system to keep the temperature in the 30–38°C range.

Dairy cattle are not the only source of manure for biogas systems. They are used as a model since dairy animals are more likely to be located in a centralized system, making the collecting and adding the manure to a digestion system less time consuming and energy intensive than for range-fed steers, or even for draft animals. Efficiencies of conversion vary not only from system to system, but also the sources of manure. Swine and beef cattle manure

TABLE V Energy Inputs Using Anaerobic Digestion for Biogas Production from 100 t wet (13 t dry) using Cattle Manure (Pimentel et al., 1988)[a,b]

	Quantity	kcal (1,000)
Inputs		
Labor hours	20 hr	—
Electricity	2,234 kWh	5,822
Cement foundation (30-year life)	0.9 kg	2
Steel (gas collector and other equipment with 30-year life)	35 kg	725
Pumps and motors	0.5 kg	1
Truck/tractor for transport (10-year life)	10 kg	200
Fuel for transport (10-km radius)	34 l	340
Total inputs		7,090
Total biogas output		10,200

[a] The retention time in the digester is 20 days. The unit has the capacity to process 1,825 t (wet) per year. Note: the yield in biogas from 100 t is estimated at 10.2 million kilocalories. Thus, the net yield is 3.1 million kilocalories. The energy for heating the digester is cogenerated from the cooling system of the electric generator.

[b] It is assumed that anaerobic digestion of the manure takes place at 35°C with a solids retention time of 20 days. The temperature of the fresh manure is 18°C, and the average ambient temperature is 13°C. The manure is assumed to have the following characteristics: production per cow per day, 23.6 kg total; solids, 3.36 kg; and biological oxygen demand (BOD), 0.68 kg. The digester is assumed to transform 83% of the biodegradable material into gas. The biogas produced is 65% methane, and its heat of combustion is 5720 kcal/m^3 at standard conditions.

appears to yield more gas per kilogram of VS than dairy cattle manure. Poultry manure is also used, but sand and other forms of heavy grit in this dung cause pump maintenance problems and require more frequent cleaning of the digester.

Manure processed in the digester retains its fertilizer value and has the advantage of less odor. Therefore, it can be spread on fields and may be easier to pump if the initial pumping system used a cutter pump to break up stray bits of straw or long undigested fibers. Biogas systems have the advantage of being able to adjust in size according to the scale of the operation. The pollution problem associated with manure in a centralized dairy production system is the same whether or not it goes through a biogas generator.

In developing countries, such as India, the situation is different. There, a substantial percentage of the manure as dried dung is burned directly as fuel. Although burning utilizes a significantly higher percentage of the total energy in the manure, it results in a complete loss of nitrogen and loss of substantial amounts of the other valuable nutrients. Whether or not biogas is a useful energy alternative in India and other similar countries is highly problematic in spite of the higher overall energy efficiency of the conversion system.

If it is not desirable to produce electricity from the biogas, the energy data listed in Table V will change considerably. For instance, less energy will be lost in the conversion to electricity if all the energy is used directly for heating. However, compressing biogas for use in tractors involves the input of significant amounts of additional energy for "scrubbing" the biogas to remove hydrogen sulfide and water.

A. Biogas for Smallholders

The economics of biogas production in a rural area of a developing nation, like Kenya or India, illustrates that costs and benefits are complex and results mixed. The capital costs of constructing a simple biogas digester with a capacity to process 8 t (wet) of manure per 20-day retention time, or 400 kg/day, are estimated to be between $2000 and $2500 (Table VI). Such a unit would have usable life of 30 years, so the capital costs are only $80 per year.

If rural workers construct the biogas generator themselves, material costs might range from $300 to $700. At $400 for materials, without any charge for labor, the investment would be only $14 per year with the costs spread out over the life of the digester.

A digester this size in India, where cows weigh an average of between 225 to 330 kg each, would require access to manure from about 20 cows. This system would produce an estimated 2277 m^3 of biogas per year at a conversion efficiency of 25% (Table VI). The energy value of this gas totals 13.0 million kcal. Assuming $8.38 per 1 million kcal, the economic value of this much energy is $109 per year. Then if no charge is made for labor and dung and the capital cost is assumed to be only $14 per year, the net return is $95 per year. These costs are not equally applicable to Kenya where the energy replacement of biogas in terms of woodfuel saved is appropriate. Using an average of 4000 kcal/kg of woodfuel, this amount of biogas would replace 3 t of wood and since biogas is generally more efficient than wood when used for cooking, the total amount of wood replaced might be double.

TABLE VI Energy Inputs for Anaerobic Digester in the Tropics for Biogas Production using 8 t (1 t dry) of Cow Manure (Pimentel et al., 1988)[a]

	Quantity (kg)	kcal
Inputs		
Cement foundation (30-year life)	0.07	140
Steel (30-year life)	0.33	7,000
Total inputs		7,140
Total biogas output		820,000
Net return per 1 t dry manure		812,840

[a] The retention time is 20 days without a means of storing the biogas. The gas is used as delivered. The digestion takes place at 35°C. The temperature of the fresh manure is assumed to be 21°C, and the average ambient temperature is 21°C. The efficiency of the digester is 25%. The biogas produced is 65% methane and its heat of combustion is 5720 kcal/m^3.

Although the labor requirement for the described biogas generator is only 5–10 min/day, the labor input for collecting and transporting biomass for the generator may be significant. If the source for the 400 kg of manure required for the digester was, on average, 3 km from the digester, it would take 2 laborers working an 8-hr day to collect manure, feed it into the digester, and return the manure to cropland where it could be utilized as fertilizer. On a per hour basis, the laborers would have to work for 3¢ per hour for the biogas digester to have costs equal to the amount of gas produced. In some situations, especially in densely populated parts of a country, the amount of transport required will be too costly.

Although the profitability of small-scale biogas production may be low even without the charge of labor, biogas digesters have significant advantages in rural areas. The biomass can be processed and fuel energy obtained without losing the valuable nutrients (N, P, and K) present in the manure. Nitrogen and phosphorus are major limiting nutrients in tropical agriculture and these are returned to the cropland. The only loss that the processed manure

has undergone is the breakdown of the fibrous material it contains, making it a less effective agent for the control of soil erosion.

In contrast, when biomass is directly burned as a fuel, both nitrogen and other nutrients are lost to the atmosphere. The nitrogen in the biogas slurry (for the 146 t/yr amounts) would amount to about 3.7 t/yr. This has an energy value of 77 million kcal and market value of $2293. Then if the nitrogen value and the gas value combined, the return for such a system is approximately $2388. The nitrogen fertilizer value of the processed manure makes it worthwhile as a biogas source rather than burning it as a primary fuel cakes. Based on this, each laborer would receive about 60¢ per hour for his work.

The total amount of manure produced annually in the United States is about one billion tons. It would be an achievement to manage to process even half of this in biodigesters. Due to low net yield of energy, as described, even 500 million t of manure, with gas produced at 28% efficiency, would provide energy for a population of 270 million Americans of 0.0076 kW per person per year. This represents only 0.0008% of present net energy use.

B. Gasification

Biomass wood with less than 50% moisture can be heated in the presence of air and gasified. The gas produced can be used to run internal combustion engines and also used as a gas fuel and for other purposes. When used in the internal combustion engine, the gas must be cleaned thoroughly as the several chemical contaminates it contains corrode engines and reduce its efficiency.

A kilogram of air-dried biomass will produce approximately 2000 kcal of clean gas which can generate about 0.8 kWh of net power electricity. The low heating value of the gas-air mixture in a gasoline engine results in derating the engine by 30–40%. This problem can be overcome by supercharging the engine. Using the gas as a mixture in a diesel engine results in derating the engine by only 10% because of its high excess in the gas-air ratio. However, the diesel engine will require a small amount of diesel fuel for ignition.

Although gasifier units can be relatively simple for small-scale operations designed, large-scale systems are most efficient. Thus, about 11.4 kcal of woodfuel is required to produce 1 kcal of gas. If the gas is cleaned, then the net return is diminished. The input : output results in an energy return in terms of wood to gas of 1 : 0.09. The equipment for cleaning the gas is expensive and uneconomical for use in rural areas, especially in developing countries. In addition to using the produced gas for internal combustion engines, it may be utilized as feedstock for various chemical products.

C. Pyrolysis

Air-dried wood or other biomass heated in the absence of oxygen can be converted into oil, gas, and other valuable fuels. The biomass feedstock, before it is fed to the pyrolysis reactor, must be ground or shredded into smaller than 14-mesh size units. Flash pyrolysis takes place at 500°C and under high pressure (101 kPa). After processing the solid char is separated from the fluids produced in a cyclone separator. The char is then used as a heating source for the reactor.

Using dry municipal refuse, the resulting products from a kilogram of biomass are water, 10%; char, 20% (energy content is about 4500 kcal/kg); gas, 30% (energy content is 3570 kcal/m^3); and oil, 40% (energy content is 5950 kcal/kg). Other investigators have reported up to 50% oil production. This gas and oil can be reprocessed, cleaned, and utilized in internal combustion engines.

The oil and gas yield from a rapid processing pyrolysis plant is about 37% or about 2.7 kcal return per kilocalorie invested. Since the plant analyzed in the study was processing city wastes, there was no energy or economic charge for biomass material. However, if tropical dry-wood is used for pyrolysis about 5 kcal of wood is required to produce 1 kcal of oil.

The gas from a gasifier-pyrolysis reactor can be further processed to produce methanol. Methanol is useful as a liquid fuel in suitably adjusted internal combustion engines.

Employing pyrolysis in a suitably large plant to produce methanol would require at least 1250 t of dry biomass per day. Based on tropical dry-wood, about 32 kcal of wood is needed to produce 1 kcal of methanol (or 1 t of wood yields 14 l of methanol). A more recent study reports that 1 t of wood yields 370 l of methanol. In either case, more than 150,000 ha of forest would be needed to supply one plant. Biomass generally is not available in such enormous quantities from extensive forests and at acceptable prices.

If methanol from biomass was used as a substitute for oil (33 quads) in the United States, about 1000 million hectare of forest land per year would be needed to supply the raw material. This land area is much greater than the 162 million ha of United States cropland now in production. Although methanol production from biomass may be impractical because of the enormous size of the conversion plants, it is significantly more efficient than ethanol production using corn based on energy output and economic use of cropland.

D. Vegetable Oil

Processed vegetable oils from sunflower, soybean, rape, and other plants can be used in diesel engines. One major

advantage of burning vegetable oils in a diesel engine is that the exhaust smells like cooking popcorn. However, the energetics and economics of producing vegetable oils for use in diesel engines are negative.

Sunflower seeds with hulls have about 25.5% oil. The average yield of sunflower seeds is 1560 kg/ha, and in terms of oil this amounts to 216 l of vegetable oil produced per hectare. This much oil has an energy value of 1.7 million kilocalories which appears promising. However, the energy input to produce this yield of 1560 kg/ha is 2.8 million kcal. Therefore, 65% more fossil energy is used to produce a liter of vegetable oil than the energy potential of the sunflower oil.

A liter of vegetable oil sells for at least $2 whereas a liter of gasoline at the pump today sells for 40¢ per liter. There is no way that vegetable oil will be an economic alternative to liquid fuels in the future.

E. Electricity

Although most biomass will continue to be used for cooking and heating, it can be converted into electricity. With a small amount of nutrient fertilizer inputs, an average of 3 t (dry) of woody biomass can be sustainably harvested per hectare per year, although this amount of woody biomass has a gross energy yield of 13.5 million kilocalories (thermal). The net yield, however, is lower because approximately 33 l of diesel fuel per hectare is expended for cutting and collecting wood for transport. This assumes an 80-km roundtrip between the forest and the electric plant. The economic benefits of biomass are maximized when the biomass is close to the processing plant.

In addition, a small amount of nitrogen fertilizer has to be applied. For bolewood, 1 t contains about 15 kg of N. Thus about 837,000 kcal is required for 3 t of bolewood.

The energy input : output ratio for the system is calculated to be 1 : 6. The cost of producing a kilowatt of electricity from woody biomass ranges from 7–10¢. This is competitive with other electricity production systems that presently have an average cost of 6.9¢ with a range of 5–13¢ per kWh. Approximately 3 kcal of thermal energy is expended to produce 1 kcal of electricity.

Woody biomass could supply the nation with about 5 quads of its total gross energy supply by the year 2050 with the use of approximately 112 million hectare (an area larger than the state of Texas). A city of 100,000 people using the biomass from a sustainable forest (3 t/ha) for fuel would require approximately 220,000 ha of forest area, based on an average electrical demand of 1 billion kilowatthours (860 kcal = 1 kWh). More than 70% of the heat energy produced from burning biomass is lost in its conversion into electricity; this is similar to losses experienced in coal-fired plants. The forest area required to supply this amount of electricity is about the same as that required to supply food, housing, industry, and roadways for a population of 100,000 people.

There are several factors that limit reliance on woody biomass. Some have proposed culturing fast-growing trees in a plantation system located on prime land. These yields of woody biomass would be higher than the average of 3 t/ha and with large amounts of fertilizers and freshwater yields might be as high as 15 t/ha. However, this is unrealistic because this land is needed for food production. Furthermore, such intensely managed systems require additional fossil fuel inputs for heavy machinery, fertilizers, and pesticides, thereby diminishing the net energy available. In addition energy is not the highest priority use of forest wood, but rather for lumber for building and pulp.

The conversion of natural forests into plantations will increase soil erosion and water runoff. Continuous soil erosion and degradation will ultimately reduce the overall productivity of the land. If natural forests are managed for maximal biomass energy production, loss of biodiversity can be expected. However, despite serious limitations of plantations, biomass production could be increased using agroforestry technologies designed to protect soil and conserve biodiversity.

IV. BIOMASS AND THE ENVIRONMENT

The presence of biomass on the land protects not only the land it covers, but also the natural interactions among all species that inhabit the ecosystem. Conversely, the removal of biomass for all purposes, but most especially for energy production, threatens the integrity of the entire natural ecosystem.

A. Soil Erosion

Once the biomass vegetation has been removed from the land area and the land is exposed to wind and rainfall energy, erosion is a major threat. Land degradation by soil erosion is of particular concern to agriculturists and foresters because the productivity of the soil is diminished. Too often soil erosion and the resulting degradation goes unnoticed (note, 1 mm of soil weighs 15 t/ha). Soil reformation is exceedingly slow. Under agricultural conditions, approximately 500 years (range from 200 to 1000 years) are required to renew 2.5 cm (340 t) of topsoil. This soil formation rate is the equivalent of about 1 t/ha/yr. Forest soil re-formation is slower than in agriculture and is estimated to take more than 1000 years to produce 2.5 cm of soil. The adverse effect of soil erosion is the gradual loss of productivity and eventually the abandonment of the land for crop production.

Serious soil erosion occurs on most of the world's agriculture, including the United States where erosion on cropland averages 13 t/ha/yr. In developing countries, soil erosion is approximately 30 t/ha/yr. The rates of erosion are intensifying in developing countries because of inefficient farming practices and because large quantities of biomass are removed from the land for cooking and heating. Rural people who are short of affordable fuels are now being forced to remove crop residues and utilize dung for cooking, leaving their soils unprotected and susceptible to wind and water erosion.

Indeed soil erosion caused by wind and water is responsible for the loss of about 30% of the world cropland during the past 40 years. For example, the rate of soil loss in Africa has increased 20-fold during the past 30 years. Wind erosion is now so serious in China that Chinese soil can be detected in the Hawaiian atmosphere during the Chinese spring planting period. Similarly, soil eroded by wind is carried from Africa to Florida and Brazil.

Erosion diminishes crop productivity by reducing the water-holding capacity of the soil and reduces water availability to the plants. In addition, soil nutrient levels and organic matter are carried away with the eroding soil and soil depth is lessened. Estimates are that the continuing degradation of agricultural land will depress world food production from 15–30% by the year 2020. Others project that Africa will be able to feed only 40% of its population in 2025 both because of population growth and soil infertility in vital cropland areas.

B. Forest Land Erosion

Forestlands lose significant quantities of soil, water, and soil nutrients wherever trees are cut and harvested. For instance, the surface water runoff from a forested watershed after a storm averaged 2.7% of the precipitation, but after forest cutting and/or farming water runoff rose to 4.5 percent. In addition, soil nitrogen leached after forest removal was 6 to 9 times greater than in forests with normal cover.

Also, the procedures used in harvesting timber and pulpwood biomass contribute to increased erosion because they expose the soil to wind and rainfall energy. Typically, tractor roads and skid trails severely disturb 20–40% of the soil surface in forests. In addition, the heavy equipment needed to harvest and clear the land compacts the soil, resulting in greater water runoff.

For example, compaction by tractor skidders harvesting Ponderosa pine reduced growth in pine seedlings from 6 to 12% over a 16-year period. Following clearing, water percolation in the wheel-rutted soils was reduced for 12 years and in log-skid trails for 8 years. This resulted in a lack of water for the remaining vegetation and limits continual forest biomass production.

C. Nutrient Losses and Water Pollution

Rapid water runoff and nutrient losses occur when crop biomass residues are harvested for fuel and rainfall easily erodes soils. Water quickly runs off unprotected soil because raindrops free small soil particles that, in turn, clog holes in the soil and reduce water infiltration. This water runoff transports soil organic matter, nutrients, sediments, and pesticides to rivers and lakes where it harms natural aquatic species. For example, conventional corn production lost an average of about 20 t/ha/yr of soil compared with only about 5 t/ha/yr with ridge- and no-till.

As mentioned, the water-holding capacity and nutrient levels of soils are lessened when erosion occurs. With conventional corn production, erosion reduced the volume of moisture in the soil by about 50% compared with no-till corn culture. In contrast, soil moisture volume increased when corn was grown in combination with living mulches. Estimates are that about $20 billion in fertilizer nutrients are lost annually from United States agriculture because of soil erosion.

Large quantities of nutrients are also lost when fuelwood and crop residues are also removed and then burned. On average, crop residues contain about 1% nitrogen, 0.2% phosphorus, and 1.2% potassium. When burned, the nitrogen is released into the atmosphere. Although some phosphorus and potassium are retained in the ashes, an estimated 70–80% of these nutrients is lost when the fine particulate matter is dispersed into the air during burning process. Thus, only a small percentage of the nutrients in crop residues are conserved even when returning the ash residues to the cropland.

D. Water Use

All biomass vegetation requires and transpires massive amounts of water during the growing season. Agriculture uses more water than any other human activity on the planet. Currently, 65% of the water removed from all sources worldwide is used solely for irrigation. Of this amount, about two-thirds is consumed by plant life (nonrecoverable). For example, a corn crop that produces about 8000 kg/ha of grain uses more than 5 million liters per hectare of water during its growing season. To supply this much water to the crop, approximately 1000 mm of rainfall per hectare, or 10 million l of irrigation, is required during the growing season.

The minimum amount of water required per capita for food production is about 400,000 l/yr. If the water

requirements for biomass energy production were added to this, the amount of required water would be more than double to about 1 million l/yr.

In addition to the unpredictable rainfall, the greatest threat to maintaining adequate fresh water supplies is depletion of the surface and groundwater resources that are used to supply the needs of the rapidly growing human population. Aquifers are being mined faster than the natural recharge rate and surface water is also not always managed effectively, resulting in water shortages and pollution that threaten humans and the aquatic biota that depend on them. The Colorado River, for example, is used so heavily by Colorado, California, Arizona, other states, and Mexico, it is usually no more than a trickle running into the Sea of Cortes.

E. Air Pollution

The smoke produced when fuelwood and crop residues are burned is a pollution hazard because of the nitrogen, particulates, and other polluants in the smoke. A report indicated that although only 2% of the United States heating energy comes from wood, and about 15% of the air pollution in the United States is caused by burning wood. Emissions from wood and crop-residue burning are a threat to public health because of the highly respirable nature of the 200 chemicals that the emissions contain. Of special concern are the relatively high concentrations of potentially carcinogenic polycyclic organic compounds and particulates. Sulfur and nitrogen oxides, carbon monoxide, and aldehydes are also released, but with wood there are usually smaller quantities than with coal.

V. SOCIAL AND ECONOMIC IMPACTS

In the future, if the world biomass is used as a major source of the world energy supply, shifts in employment and increases in occupational health and safety problems can be expected. Total employment would be projected to increase 5% if about 11% of the United States energy needs were provided by biomass. This labor force would be needed in agricultural and forest production to plant, cut, harvest, and transport biomass resources and in the operation of various energy conversion facilities.

The direct labor inputs for wood biomass resources are 2–30 times greater per million kilocalorie than coal. In addition, a wood-fired steam plant requires 2–5 times more construction workers and 3–7 times more plant maintenance and operation workers than a coal-fired plant. Including the labor required to produce corn, about 18 times more labor is required to produce a million kilocalories of ethanol than an equivalent amount of gasoline.

Associated with the possibilities of increased employment are greater occupational hazards. Significantly more occupational injuries and illnesses are associated with biomass production in agriculture and forestry than with either coal (underground mining), oil, or natural gas recovery operations. Agriculture and forestry report 61% more occupational injury and illness rates than mining. In terms of a million kilocalories of output, forest biomass has 14 times more occupational injuries and illnesses than underground coal mining and 28 times more than oil and gas extraction. Clearly, unless safe harvesting practices and equipment are developed and used, increased forest harvesting and agricultural production for energy will result in high levels of occupational injuries and increased medical expenditures and workman compensation.

The future development of major biomass energy programs will require large amounts of cropland suitable for biomass production and ultimately result in increased prices for some consumer commodities. The use of commodities, especially grains, for energy leads to competition with traditional uses of these commodities. Thus, with increased grain use for ethanol production, inflation of farm commodity prices could result. This in turn would increase farmland prices and make it more difficult for new farmers to enter the business and for existing small farmers to cope with higher rents, taxes, interest payments, and production costs. Food prices in supermarkets would be expected to increase.

VI. CONCLUSION

Certainly increased use of biomass as a fuel could provide the United States and the world with more renewable energy. A major limitation of biomass energy production includes the relatively small percentage (average 0.1%) of light energy that is captured by the earth's plant material. This governs how much biomass can be produced per unit land area. In addition to solar energy, suitably warm temperature conditions, adequate amounts of water, and the absence of pests are essential for plant growth. In North America, for example, plant growth only occurs for approximately three months of the year. In arid regions of the world plant growth is restricted only to periods of adequate rainfall.

The removal of biomass, such as crop residues, from the land for energy production intensifies soil erosion, water runoff, and soil nutrient losses. In addition, the conversion of natural ecosystems into energy-crop plantations would alter and/or reduce the habitat and food sources for wildlife and biodiversity.

At present, about half of the world's biomass is harvested as food and forest products. Thus, there is a limit

as to how much biomass can be harvested as an energy source without further causing the extinction of more plants, animals, and microbes because of biomass resources on which biodiversity depends. Agriculture and managed forests occupy approximately 70% of the total land area and use about 70% of the total water consumed by society, and this further limits natural biodiversity.

However, opportunities do exist to combine agriculture and forest production. If this is to be done several changes would have to be made in many technologies now used in agriculture and forestry. These technologies include conserving soil, water, and nutrient resources. Of particular importance is keeping the land covered with vegetation and maintaining high levels of organic matter in the soil.

Although biomass resources have a lower sulfur content than oil and coal, biomass energy conversion and use has associated environmental and public health problems. For example, the chemical emissions from wood-burning for cooking and heating produce serious chemical pollutants, including some carcinogens and other toxicants. In addition, on the basis of a million kilocalorie output, harvesting forest biomass energy is about 14 times more hazardous than coal and oil mining.

Ethanol production using grains and other food material for gasohol can be expected to have a significant negative impact on social and economic systems. A major ethanol program would help fuel inflation by raising food prices to the consumer. In addition, "burning food" as ethanol in automobiles has serious political and ethical considerations.

In conclusion, the conversion of biomass to provide an energy source has some potential to contribute to world energy needs, but the associated environmental, health, social, and economic problems must be carefully assessed. The foremost priority is the supply of food. Especially vital to this goal is maintaining an ample supply of fertile cropland needed to feed the rapidly growing world population.

ACKNOWLEDGMENT

I sincerely thank the following people for reading an earlier draft of this article and for their many helpful suggestions: Andrew R. B. Ferguson, Optimum Population Trust, U.K.; Marcia Pimentel, Division of Natural Sciences, Cornell University; Joel Snow, Iowa State University; and Paul Weisz, Pennsylvania State University.

SEE ALSO THE FOLLOWING ARTICLES

Bioreactors • Energy Flows in Ecology and in the Economy • Greenhouse Effect and Climate Data • Pollution, Air • Pollution Control • Renewable Energy from Biomass • Waste-to-Energy Systems • Water Pollution

BIBLIOGRAPHY

Ellington, R. T., Meo, M., and El-Sayed, D. A. (1993). "The net greenhouse warming forcing of methanol produced from biomass," *Biomass Bioenergy* **4**(6): 405–418.

Ferguson, A. R. B. (2000). "Biomass and Energy," The Optimum Population Trust, Manchester, U.K.

Pimentel, D. (1991). "Ethanol fuels: Energy security, economics, and the environment," *J. Agr. Environ. Ethics* **4,** 1–13.

Pimentel, D., Doughty, R., Carothers, C., Lamberson, S., Bora, N., and Lee, K. "Energy inputs in crop production in developing and developed countries," *J. Agr. Environ. Ethics*, in press.

Pimentel, D., and Kounang, N. (1998). "Ecology of soil erosion in ecosystems," *Ecosystems* **1,** 416–426.

Pimentel, D., and Krummel, J. (1987). "Biomass energy and soil erosion: Assessment of resource costs," *Biomass* **14,** 15–38.

Pimentel, D., and Pimentel, M. (1996). "Food, Energy and Society," Colorado University Press, Boulder, Colorado.

Pimentel, D., and Strickland, E. L. (1999). "Decreased rates of alluvial sediment storage in the Coon Creek Rasin, Wisconsin, 1975–93," *Science* **286,** 1477–1478.

Pimentel, D., Rodrigues, G., Wang, T., Abrams, R., Goldberg, K., Staecker, H., Ma, E., Brueckner, L., Trovato, L., Chow, C., Govindarajulu, U., and Boerke, S. (1994). "Renewable energy: economic and environmental issues," *BioScience* **44,** 536–547.

Pimentel, D., Warneke, A. F., Teel, W. S., Schwab, K. A., Simox, N. J., Ebert, D. M., Baenisch, K. D., and Aaron, M. R. (1988). "Food versus biomass fuel: Socioeconomic and environmental impacts in the United States, Brazil, India, and Kenya," *Adv. Food Res.* **32,** 185–238.

Shapouri, H., Duffield, J. A., and Graboski, M. S. (1995). "Estimating the Net Energy Balance of Corn Ethanol," Agricultural Economic Report, Washington, DC.

Tripathi, R. S., and Sah., V. K. (2000). A biophysical analysis of material, labour and energy flows in different hill farming systems of Garhwal Himalaya, "Agriculture, Ecosystems and Environment," in press.

WHO (1996). "Micronutrient Malnutrition—Half of the World's Population Affected," No. 78, 1–4, World Health Organization.

Biomass, Bioengineering of

Bruce E. Dale
Michigan State University

I. Background
II. Characteristics of Biomass
III. Uses of Biomass
IV. Bioprocessing of Biomass
V. Potential and Limitations of Biomass and Biobased Industrial Products

GLOSSARY

Biomass Plant material.

Bioprocessing Any chemical, thermal, physical or biological processing done to biomass to increase its value.

Biobased industrial products Plant-derived chemicals, fuels, lubricants, adhesives, plastics—any and all industrial products derived from biomass that are not used for human food or animal feed. For purposes of this article, *biomass is bioprocessed into biobased industrial products.*

Biorefineries Large, highly integrated facilities, analogous to petroleum refineries, that process biomass to biobased industrial products and other value-added products.

Life cycle analyses Comprehensive inventories of the material and energy flows required to produce, use and dispose of specific products throughout their entire life cycles.

Lignocellulose The structural portion of most plants, composed of a complex mixture of cellulose, hemicellulose and lignin and comprising the vast majority of all biomass. Cellulose is a polymer of glucose (sugar) while hemicellulose is a polymer made up of a variety of sugars. Lignin is a complex polymer of phenylpropane units.

Sustainable development Economic development that meets the legitimate needs of current generations without compromising the ability of future generations to meet their own needs.

BIOMASS is the only potentially renewable source of organic chemicals, organic materials and liquid transportation fuels. The biomass resource is huge. While estimates are necessarily imprecise, it is believed that photosynthesis fixes approximately 150 billion tons of new plant matter annually on the planet. Production of biobased industrial products has the potential to benefit both the economy and the environment and to provide new pathways for sustainable economic development.

The energy value of our renewable plant resource is approximately ten times the total energy value of all other forms of energy used by humanity including all fossil fuels, hydropower, nuclear energy and so on. Biomass is also relatively inexpensive and compares favorably with petroleum on a cost per pound basis and, frequently, on

a cost per unit of energy basis. Although the biomass resource is huge and comparatively inexpensive, we have invested much less effort in learning how to bioprocess or convert it efficiently to biobased industrial products than we have invested in converting petroleum to meet our needs for fuels, chemicals, materials, and other industrial products.

Compared to the petroleum processing industry, the biomass processing industry is still relatively underdeveloped, although the biomass processing industry is in fact already very large and is also growing rapidly. Thus much of this article deals with what is required for the biomass processing industry to grow further and what some of the possible and desirable growth paths for this industry might be.

I. BACKGROUND

The potential benefits (including economic, environmental and national security benefits) of obtaining a larger fraction of our fuel and chemical needs from biomass rather than from petroleum have driven increasing interest in biobased industrial products in the United States and many other countries. Lack of cost–effective bioprocessing technology is perhaps the principal barrier to more economical production of biobased industrial products. Although biomass is abundant and low cost, unless we learn how to cost-effectively convert biomass to these industrial products, their potential benefits will be largely unrealized.

While the potential benefits of biobased products are certainly real, it is also correct that unless such products are produced with proper intelligence and care, their benefits may be reduced or even negated. We must be careful that biomass is grown, harvested, converted to industrial products, and that these products are used and disposed of, in sustainable, environmentally sound systems. Careful, thorough and easily verified life cycle analyses will help us realize the potential of biobased industrial products to benefit our economy and our environment and also to avoid potential problems with the production and use of these products.

One of the most important areas demanding careful life cycle (whole system) attention for biomass conversion to industrial products is the potential conflict with food and feed production. Biomass production for biobased industrial products seems to conflict with use of the same agricultural resources for human food and animal feed. This article briefly addresses this crucial point and finds considerable room for optimism.

II. CHARACTERISTICS OF BIOMASS

A. Production of Biomass

1. Natural Inputs to Biomass Production

Natural (or ecosystem) inputs to biomass production are soil (including the associated nutrients and living organisms found in soil), genetic information, air, water, and sunlight. All of these inputs are potentially renewable indefinitely with proper oversight and intelligent design. In fact, biomass production has the potential to improve soil, water, and air quality. The entire life cycle of biomass production, bioprocessing, and biobased product use and disposal should be examined carefully to discover and properly exploit such opportunities. Intelligent design of biomass processing systems should take advantage of opportunities to improve the environment and enhance ecosystem stability under circumstances peculiar to each region and product. With careful and thoughtful design, biomass production and processing can increase or enhance the "natural capital" of soil, air, and clean water upon which all life depends.

Human inputs to biomass production include additional plant nutrients beyond those provided through the ecosystem, plant genetic improvement, labor, financial capital and intelligence, as referred to above. Much agriculture is also practiced with large inputs of fossil fuels. As mentioned, thorough and careful life cycle analysis is required to determine whether biomass processing to biobased products actually fulfils its potential to give us a more sustainable economy.

2. Potential and Actual Yields of Biomass

A key factor determining the economic (and therefore the resulting ecological) benefits of biomass production and processing is the yield of biomass, defined as the annual production of biomass (dry weight) per unit land area, often expressed as tons of dry biomass per acre per year. Meeting legitimate human needs by more intensively producing biomass (i.e., increasing yields) will allow larger tracts of land to be set aside for recreation, parks, and biological reserves. Biomass yields vary widely. The upper limit of solar energy conversion efficiency by biomass appears to be about 12% (incoming solar energy converted to the energy content of plant material). Yield seems to be tied closely to conversion efficiency; the higher the conversion efficiency, the higher the yield. Sugarcane is one of the more efficient crops, with solar energy capture efficiencies in the neighborhood of 2 to 3% and corresponding biomass yields of between 25 and 35 dry tons per acre per year. The corresponding efficiency value for corn is about 0.8%.

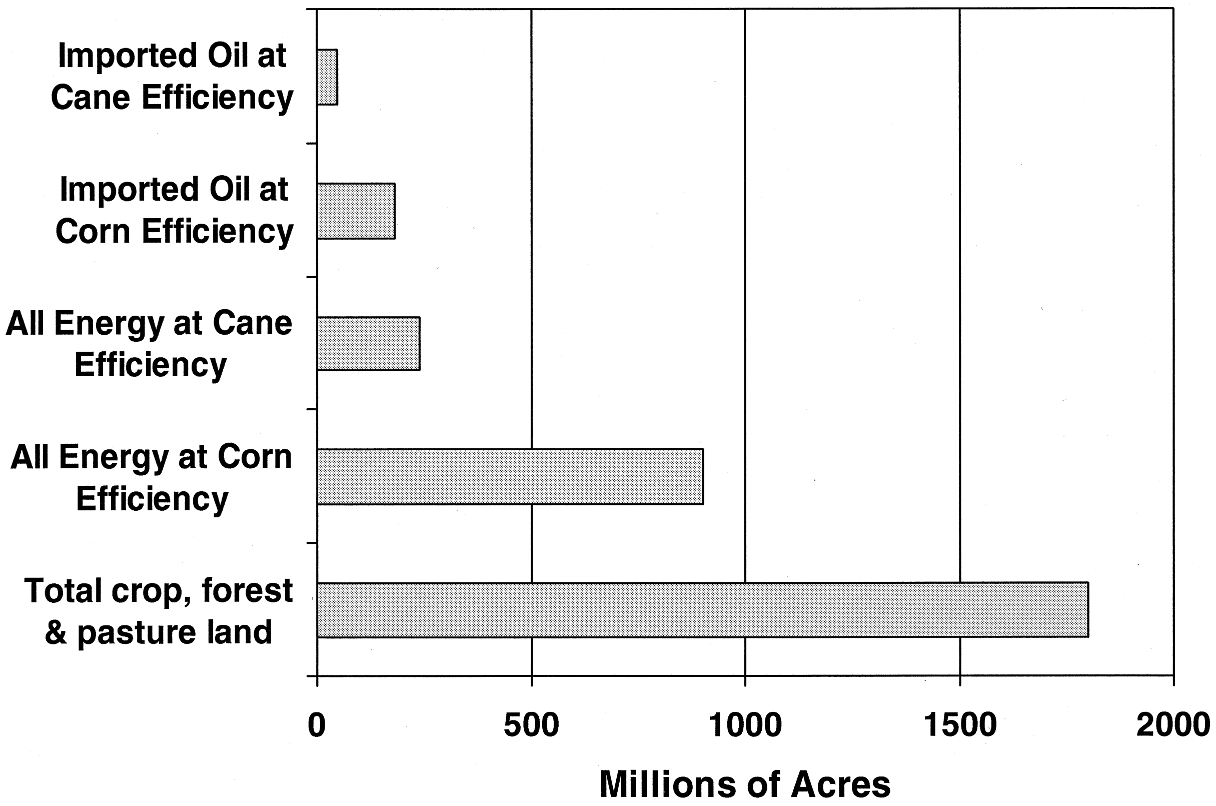


FIGURE 1 U.S. land required for biomass energy.

Increasing biomass yields is a crucial area for research. We have invested much effort and money in increasing the yields of grains such as corn. Average per acre corn yields increased at a rate of over 3% per year between 1950 and the present: corn yields were about 28 bushels per acre per year in 1947 and topped 127 bushels per acre per year in 1997. However, we have done comparatively little genetic or agronomic research to increase the yields of the perennial grasses and tree crops on which a sustainable biomass processing industry will likely be based. Thus there is great room for improving these yields.

Biomass is currently the most practical collector we have of solar energy on a large scale. The solar energy incident on the United States each year is about 600 times our annual energy consumption of about 95 quads (one quad equals one quadrillion BTU or one million billion BTU). The higher the biomass yields, the more solar energy is collected per unit land area. At a solar energy conversion efficiency of 0.8% (corn efficiency), approximately 40% of the U.S. land area placed in continuous crop production would produce biomass with an energy value equal to our total use of energy from all forms. At this efficiency, about 10% of our land area, or 100 million acres, would be required to produce the energy equivalent of all of the petroleum we use. This is roughly equal to the land currently in hay production (60 million acres) plus land idled under government programs. Obviously, other inputs in addition to solar energy are required for biomass production. Nonetheless, these statistics gives some idea of the potential to meet our energy needs from biomass. Figure 1 summarizes some of figures for U.S. land area usage and the area required to equal our energy usage at solar energy conversion efficiencies typical of corn and sugar cane.

3. Comparison of Biomass and Petroleum

Worldwide consumption of crude oil, a large but nonetheless limited resource, was about 27 billion barrels per year in 1999 or about 4 billion tons, with an approximately equal amount of coal consumed. As mentioned earlier, total production of new biomass, an indefinitely renewable resource, is approximately 150 billion tons per year.

The energy value (heat of combustion) of petroleum is about twice that of biomass (trees have a higher energy content than grasses) while coal averages about one and a half times the energy value of biomass. The lower energy value of biomass is due to the fact that it contains substantial oxygen, while petroleum has little or no oxygen.

The lower energy content (i.e., the higher oxygen content) of biomass is both an advantage and a disadvantage for this renewable resource. Biomass and the many oxygenated compounds that can be made from biomass are inherently more biodegradable and therefore more environmentally compatible than petroleum and petroleum-derived compounds. Put another way, a large spill of wheat straw is not an environmental disaster, while we are well aware of the impacts of petroleum spills. Powerful economic considerations tied to raw material use efficiency also direct us toward maintaining the oxygen molecules in biobased industrial products.

Petroleum, a liquid, is easier and less expensive to transport and store than solid biomass. One consequence of this fact is that much biomass processing will likely be done relatively close to where the biomass is produced. This may provide opportunities to integrate biomass production with biomass processing and to more easily return to the land the unused or unusable components of biomass. Few such opportunities exist with petroleum processing. In many climates, biomass production takes place only during part of the year, so there are additional storage issues that are unique to biomass. Finally, large quantities of biomass can be produced in many, if not most, countries, while comparatively few countries produce significant quantities of petroleum. Thus biomass production is inherently more "democratic" than petroleum production and is certainly less susceptible to political manipulation.

4. Cost of Biomass versus Fossil Feedstocks

Petroleum costs varied between about $10 and $20 per barrel ($65 to $130 per ton) during the decade of the 1990s. Currently oil prices are about $30 per barrel or roughly $200 per ton. Coal is available for approximately $30 per ton. By comparison, corn at $2.50 per bushel, an "average" corn price over the last decade, is roughly equivalent to $90 per ton. Corn is currently less than $2.00 per bushel, or about $70 per ton, approximately one third the current price of crude oil. Hay crops of different types and qualities are available in very large quantities (tens of millions of tons) for approximately $30–$50 per ton and several million tons per year at least of crop residues such as rice straw and corn stover are available in the United States for less than $20 per ton. Figure 2 summarizes some of these comparisons of the relative prices of biomass and fossil resources. Worldwide, many hundreds of millions of tons of crop residues such as rice straw, sugar cane bagasse and corn stover are likely to be available at very low cost, probably less than $20 per ton. Thus while fossil resources are relatively inexpensive (even given oil price volatility) renewable plant resources are equally inexpensive, and in many cases, less expensive. The importance of this fact to biomass processing cannot be overstated.

Plant raw material costs are crucial for the development of cost-competitive biobased products. For well-developed processes making commodity chemicals and fuels, approximately 50–70% of the total production costs are due to the raw material costs. Thus inexpensive biomass should eventually lead to inexpensive biobased products, *if the necessary bioprocessing technologies for converting biomass to biobased products can also be made inexpensive.* In general, we do not yet have inexpensive biomass processing technology. However, if the necessary research and development work is done to learn how to inexpensively convert biomass to biobased products, there is every reason to believe that these biobased products can compete on a cost and performance basis with similar products derived from petroleum.

To illustrate, the large chemical companies Dow Chemical and DuPont have recently announced plans to produce monomers for polymer (plastic) production from renewable sugars and starches. These carbohydrates are relatively inexpensive, and the companies have also developed inexpensive and effective conversion technologies to produce the monomers. For example, Cargill Dow Polymers (CDP) LLP (Minnetonka, MN) is building the first of up to five large processing plants to convert corn starch into lactic acid and then into polymers (polylactides). Although biodegradability of the polymers is obviously a benefit, CDP expects its polylactide polymers to compete on a cost and performance basis with petroleum-derived competing products. Similarly, DuPont's carbohydrate-derived product, 1,3 propanediol, is intended to compete directly with the identical molecule produced from petroleum. The chemical industry is therefore beginning to change its raw material base. As technologies improve and bioprocessing costs decrease, there is every reason to believe that more such products will follow.

B. Major Types of Biomass: Their Production and Composition

1. Sugar Crops

The major sugar crops are sugar cane and sugar beets. Worldwide, approximately 100 million tons per year of sugar (sucrose) are produced from sugar cane and sugar beets. Most of these sugars are used ultimately in human

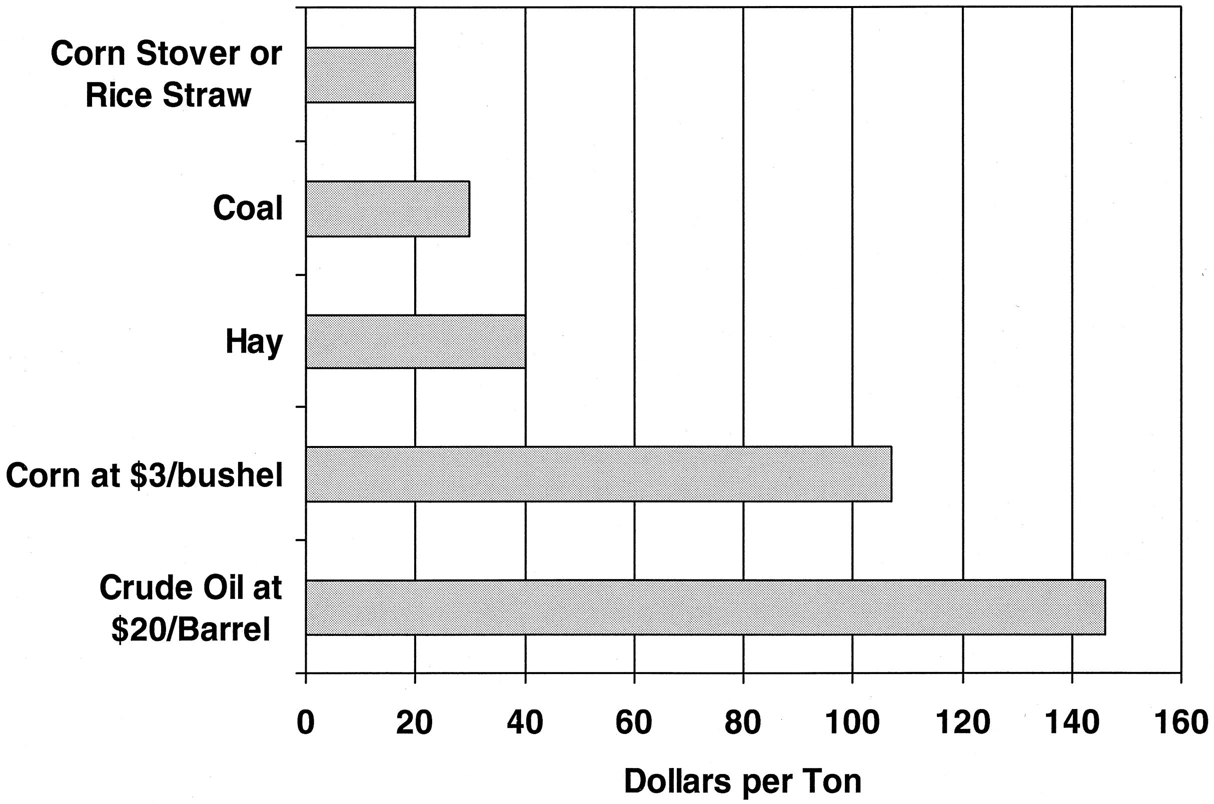


FIGURE 2 Cost of fossil vs. biomass feedstocks.

nutrition. Sugar cane is grown largely in the tropics while sugar beets are grown mostly in temperate zones.

Approximately 10 dry tons of a fibrous residue (called bagasse) are produced for every ton of cane sugar while about 0.5 dry tons of residue are produced for every ton of beet sugar. Worldwide the total production of sugarcane bagasse is approximately 800 million metric tons per year. These residues have few other uses and represent a very large potential resource for bioprocessing to biobased industrial products.

Sugar cane bagasse consists of approximately 40% cellulose, 30% hemicellulose and 20% lignin with small amounts of minerals, sugars, proteins and other compounds. The composition of bagasse is, however, variable depending on growing conditions, harvesting practices and processing methods. Beet sugar residue consists of approximately equal portions of cellulose, hemicellulose and pectins, with a few percent of lignin. Cellulose and hemicellulose are polymers of glucose and other sugars. However, the sugars in cellulose, and to a lesser degree those in hemicellulose, are not very good animal feeds and they are essentially indigestible as human foods. Microbial and enzymatic conversion of these sugars in hemicellulose and cellulose to biobased products is also significantly limited for the same reasons cellulose and hemicellulose are not easily digested by animals. The resistance of cellulose to conversion to simple sugars is a key limitation in biomass processing that must be overcome if biomass processing is to attain its full potential. To put the potential of cellulose and hemicellulose in perspective, the potential sugar that might be produced from cellulose and hemicellulose in sugar cane bagasse alone is approximately six times the total sugar produced worldwide by both sugar cane and sugar beets.

2. Starch Crops

A wide variety of starch crops (mostly grains) are grown worldwide, including corn (maize), rice, wheat, manioc (cassava), barley, rye, potatoes and many more. At least 2 billion tons per year of these crops are produced worldwide. While most sugar is used to feed human beings, much grain is used to feed animals, particularly in the more developed countries. One key indicator of a country's

economic growth, in fact, is a substitution of meat for grains in the human diet. Animals convert grain to meat or milk with widely varying efficiencies, however. Fish, poultry, and swine are relatively efficient converters, while beef cattle are considerably less efficient.

Most grain crops produce a byproduct, or residue, that is primarily composed of cellulose, hemicellulose, and lignin, called collectively lignocellulose. Thus very large tonnages of rice straw, corn straw (called corn stover), and many other straws are produced as a low value (and low cost) byproduct of grain production. Approximately 1 to 2 tons (dry weight) of these straws are produced per dry ton of grain. Using this ”rule of thumb,” the total worldwide production of just corn stover and rice straw is in the neighborhood of 1 billion tons per year. Taken together with sugar cane bagasse production, the total amount of corn stover, rice straw and bagasse produced each year is approximately 2 billion tons.

Very large quantities of other straws and crop processing residues are also produced. Many such residues are produced at centralized processing facilities. While some of this residual plant matter should be left on the field, much of it can be removed and used elsewhere without degrading soil fertility. For example, rice straw is often burned to clear the fields for the next crop. There is considerable political pressure in the United States and elsewhere to eliminate or greatly reduce this practice of field burning.

3. Plant Oil and Protein Crops

There are many different plant oil crops including soybeans, palm, coconut, canola, sunflower, peanut, olive and others. The total worldwide production of fats and oils by these crops exceeds 75 million tons per year, with an additional 12 million tons per year or so of animal fats. (An oil is simply a liquid fat.) Most plant oils go into human foods or animal feeds. However, there is a very long history of also using and modifying plant oils for fuels, lubricants, soaps, paints and other industrial uses. Oils consist chemically of an alcohol (glycerol) to which are attached three long chain carboxylic acids of varying composition. Plant oil composition varies widely with species and the composition strongly affects the industrial uses to which these oils can be put. Therefore by modifying these oils, they can potentially be tailored to desired applications.

The other major product of oilseed crops is a high protein “meal,” usually produced by expelling or extracting the oil from the seed. Total world production of high protein meals from oilseeds is approximately 180 million tons per year. The predominant oilseed meal is soybean meal containing approximately 44% protein. While there are some industrial uses for this meal, the bulk of it is fed to animals.

As with the starch crops, most of these oilseed crops produce one or more residues that are rich in lignocellulose. For example, soybean straw is typically left in the fields when the beans are harvested. Soybean hulls are produced as “wastes” at the oilseed processing plant. In the United States, approximately 10 million tons per year of these soybean hulls are produced as a byproduct of soybean crushing operations.

4. Tree and Fiber Crops

In contrast with the crops mentioned, essentially all of the wood harvested is destined for industrial uses, rather than food/feed uses. Production of wood for lumber in the United States amounts to about 170 million tons per year while U.S. pulpwood production (destined for all kinds of paper uses) is about 90 million tons/year. A wide variety of industrial chemicals such as turpentine, gums, fats, oils, and fatty acids are produced as byproducts of pulp manufacture.

Not all paper is derived from trees, however. Some grasses and crop residues such as kenaf and sugar cane bagasse have been used or are being considered as fiber/paper crops. The giant reed kenaf, in particular, has very rapid growth rates and favorable agronomic characteristics. A major impediment to its introduction as an alternative newsprint crop seems to be the huge capital investment required for a new pulp and paper plant.

The growing worldwide demand for paper products of all kinds may limit the ability to use tree and pulpwood crops for other industrial applications, given the value of long plant fibers in paper production. Even short rotation woody crops (trees grown for energy use as if they were grasses), must cope with the demand for that land and the long fibers grown on it for pulp and paper uses. Typical pulp prices are in the neighborhood of $600 per ton or $0.30 per pound, a high raw material cost hurdle indeed for commodity chemicals that are often targeted to sell for less than $0.30 per pound. Some residues from fiber crop production and processing may be available at much lower cost and could perhaps be used for chemical and fuel production. Typically these residues are burned to get rid of them and recover at least their energy value.

The most important fiber crop is cotton. Worldwide production of cotton in 1998 totaled about 91 million bales, each weighing about 480 lb. Given the high value textile uses of cotton, it is similarly unlikely that much cotton will be devoted to other uses. However, there are many millions of tons of wastes generated at cotton gins and mills that might be used industrially if appropriate, low-cost, conversion technologies were available. Chemically, these tree and fiber crops and their residues are essentially all

lignocellulosic materials, i.e., they are composed mostly of sugar polymers and lignin.

5. Forage and Grass Crops

For purposes of this article, we will not distinguish between grasses and legumes, but will consider all non-woody annual and perennial plants as "grasses." Most grasses utilized by humans are employed for animal feeding. Most forage grasses are also produced for local use and are not widely traded, making it difficult to establish a "market price" for many grasses.

Available statistics on forage and grass crops are much less complete than for the sugar, starch and oilseed crops. However, there are approximately 7 billion acres worldwide devoted to animal pasture. If we assume that only 1 ton of forage grasses is produced per acre per year on such lands (the U.S. average for managed hay lands is approximately 3 tons per acre per year), the total amount of animal forage from pasturelands is about 7 billion tons per year, on a worldwide basis. In the United States we produce at least 300 million tons per year of mixed forage grasses (dominated by alfalfa).

Forages and grasses vary widely in composition, although they can be considered lignocellulosic materials. However, grasses contain a much wider variety of components that do most tree species. In addition to cellulose, hemicellulose and lignin, grasses often contain 10% or more of protein, in addition to minerals, starch, simple sugars, vitamins and other components. The wider variety of components in grasses versus woody plants offers the potential for producing additional valuable products, but may also complicate the processing required.

To summarize this section, in very rough terms the world's agricultural system produces about 2.5 billion tons per year of total sugar, starch, oil, and plant protein to feed both humans and animals, as well as for some industrial uses. At least this much crop residue is also produced as a byproduct of sugar, starch and oilseed crops. Crop residues are generally lignocellulosic materials. Additionally, well over 10 billion tons per year of lignocellulosic materials are grown with some degree of human involvement as crop and forest residues, animal forages and pasture, not including the managed production of timber and pulpwood. Many more billions of tons of lignocellulosic materials are produced annually in the biosphere with essentially no human intervention. Thus the size of the lignocellulosic resource dwarfs that of the food/feed resource represented by the sugar, grain and oilseed crops. Figure 3 attempts to summarize these data on the annual production of biomass

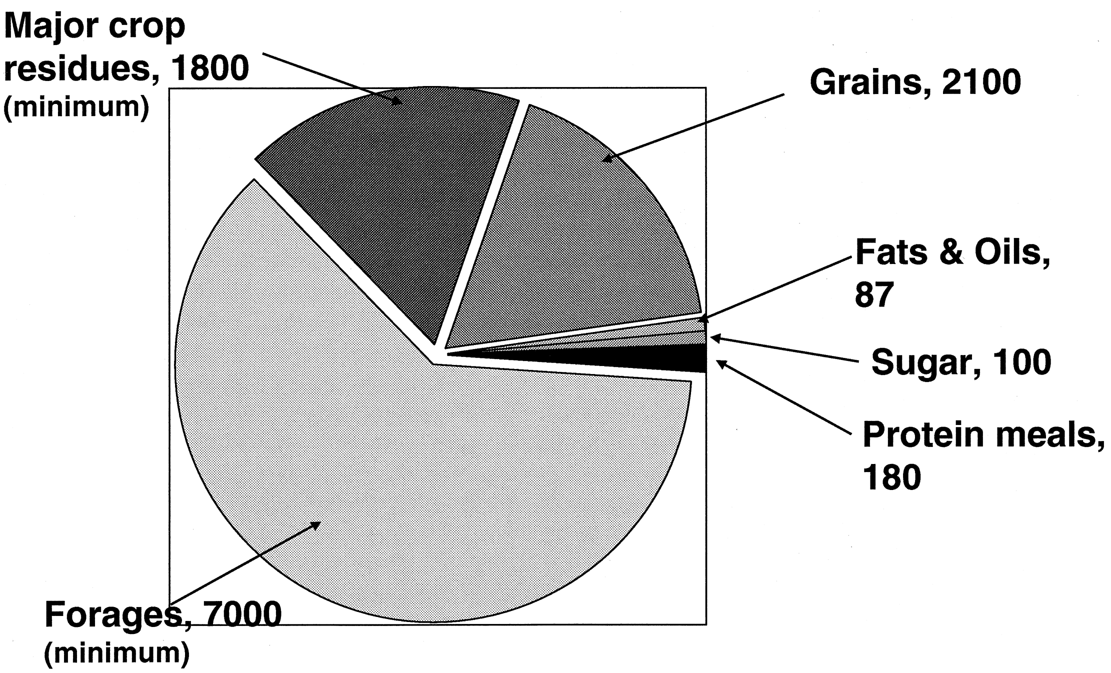


FIGURE 3 World food and forage production (millions of tons).

for these many food and feed uses, as well as available crop and forestry residues.

C. Biotechnology and Biomass Production

1. Modify Biomass Composition for Easier Processing

Plant breeding and/or molecular biology techniques can be used to alter the composition of plant matter to make it easier to process. As mentioned previously, given the already relatively low cost of biomass, it is believed that reducing the costs of converting biomass to industrial products will be the key factor in helping these products compete with petroleum-derived products on a much larger scale. For example, reducing the lignin content of grasses and trees should make it easier to convert the cellulose and hemicellulose portions to sugars that might then be fermented or otherwise processed to a wide variety of chemicals and fuels. Changing the fatty acid composition of a particular plant oil could improve the ability to separate that oil in a processing facility. The possibilities are quite literally endless. The ability to modify the raw material composition and properties has no parallel in petroleum refining (or hydrocarbon processing generally). This is a major potential advantage of biobased industrial products that should be exploited whenever possible.

2. Enhance Biomass Yields and Reduce Inputs

As mentioned, both plant breeding and molecular biology can be used to increase the yields of biomass grown for industrial uses and to reduce the inputs required to produce these industrial crops. High yields are important both to reduce the costs of biobased products and to decrease the total amount of land required to supply these products. Reductions in crop production inputs such as fertilizers, pesticides, herbicides and even water will also tend to reduce the costs of biomass production and could have very large, positive environmental effects. For example, deep-rooted perennial grass species or trees destined for conversion to industrial products might be planted around fields devoted to row crops such as corn and at the edges of streams to intercept fertilizers and pesticides in groundwater and to reduce soil erosion. Agricultural chemicals in runoff are believed to contribute significantly to oxygen-depleted, and therefore life-depleted, regions in the Gulf of Mexico and elsewhere.

3. New Products

Breeding has long been used to alter the composition of plant materials, for example to increase the content of various starch fractions in corn or to modify the sugar content of sugar cane. Plant breeding has also been used to alter the composition and amounts of various biomass fractions for industrial uses, for example, to increase the resin (rubber) content in the desert shrub called guayule. Such plant breeding efforts are relatively uncontroversial.

However, molecular biology/genetic engineering can also be used to modify the existing genes of plants and to transfer entirely new genes into plants. For example, bacterial genes for a biodegradable plastic called have been successfully expressed in plants, leading to the possibility of "chemical plants in green plants." This is an exciting and technically very promising possibility. It is also a possibility with potentially great environmental benefits. However, considerably more political and environmental controversy is likely to surround such efforts. Careful studies will be needed to demonstrate that expression of foreign genes in plants destined for industrial uses will not lead to undesired environmental or human health consequences.

III. USES OF BIOMASS

A. Current Uses

1. Food/Feed Consumption and World Protein/Calorie Demand

Average human requirements are approximately 2000 kcal of food energy and about 50 g of protein daily, in addition to smaller amounts of essential oils and vitamins. Assuming a total world population of 6 billion people, then the total human demand for protein is approximately 120 million tons/year and the total calorie requirement is about 4.5 million billion kcal/year (4.5×10^{15} kcal/year). Total world grain (corn, wheat, rice, oats, sorghum, barley, millet and mixed grains) production in 1998/1999 was approximately 2.1 billion tons.

If we assume that grain contains on average 70% carbohydrate (sugars) and 11% protein, then this grain crop alone is sufficient to supply all of the calorie needs of the world's people and about 50% more than the protein needs of all of humankind. If we include the additional calories and protein available from sugar cane and sugar beets, oilseeds and a myriad of other crops such as potatoes and cassava, the total worldwide production of calories and proteins is several fold greater than the human demand.

Obviously, much of the plant matter we grow is used to feed animals, not people directly. However, if we so chose, we could easily feed the world's population with an adequate, plant-based, diet using a fraction of the land now devoted to agriculture and animal husbandry. (There

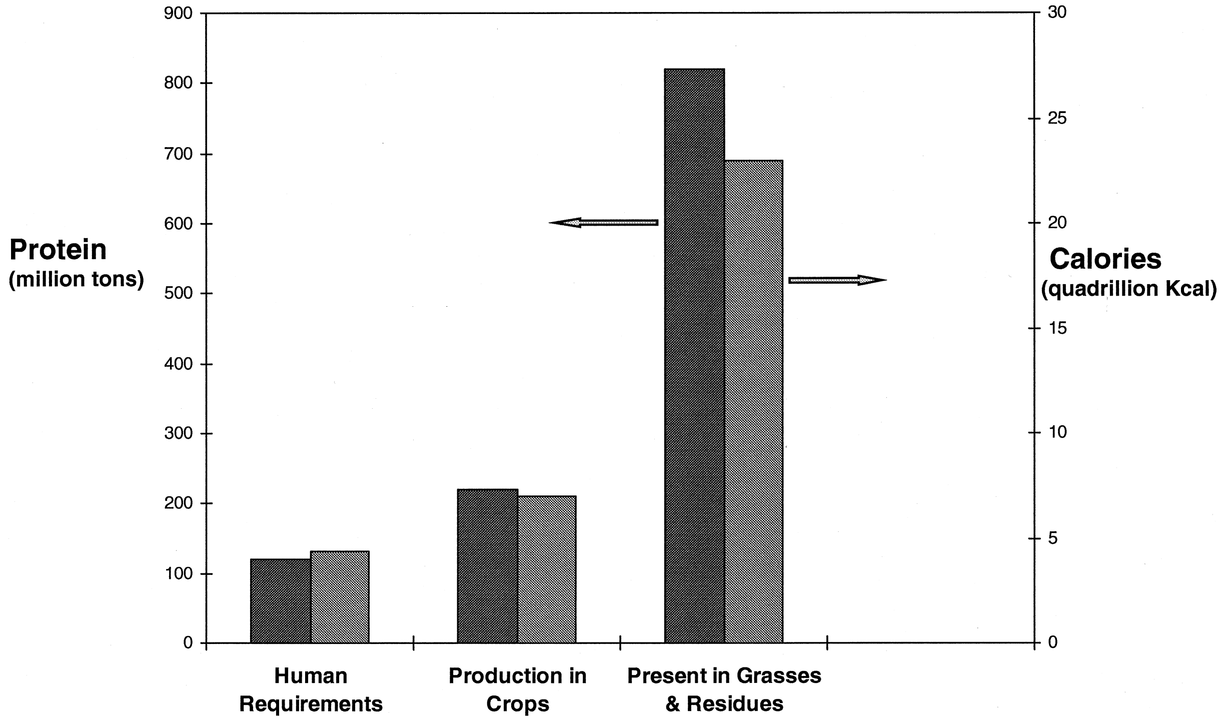


FIGURE 4 Human needs for protein and calories vs. nutrient production in crops and lignocellulosics.

is also some consumption of plant matter for industrial uses, the subject of this article.)

Past history suggests that people seek to increase their consumption of meat, milk and eggs as their income grows. Looking at these consumption figures in a different light, if we found another way to meet more of the protein and energy (calorie) needs of our animals from sources other than grains and oilseeds, we could then free up large quantities of grain and oilseed crops for other uses, including industrial uses. From the crop production statistics quoted above it is obvious that the potential and actual production of grasses and other lignocellulosic materials far exceeds the production of grains, oilseeds and sugar crops.

2. Animal Feeds

In 1998 the United States produced about 40 million tons of beef, pork, and poultry as well as billions of dozens of eggs and tens of millions of tons of milk. To generate these products, livestock and poultry consumed well over 500 million tons of feed expressed on a feeding value equivalent to corn. Over half of this total feed was from forages, about two thirds of which was grazed as pasture and the rest of which came from harvested forages such as hay and silage. The remainder of animal feed consumed was concentrates such as corn, sorghum, oats, wheat, etc. If it were possible to derive more and better animal feeds from forages and other lignocellulosic materials, it might be possible to increase the use of agricultural raw materials for biobased industrial products without negatively impacting food availability and food prices.

3. Existing Fuels/Chemicals/Materials Uses of Biomass

As described above, biomass has long been used as a solid fuel, as a building material and also as a source of fiber for clothing and paper. These uses continue. In the United States, the forest products industry is valued at $200 billion per year and the cotton produced is worth another $5 billion per year. Prior to the early 1800s, biomass was in fact the chief source of fuel and materials. With the coming of the Industrial Revolution, a gradual switch from biomass as the major fuel source took place, first through a transition to coal and later to petroleum and natural gas. The oil

refining industry was developed over about the last 120 years and catalyzed the development of a huge chemical industry, based mostly on petroleum as the ultimate raw material. Today in the United States, the chemical process industries have total sales of over $360 billion per year and the petroleum refining industry is worth about $250 billion per year.

Use of biomass for chemicals and materials is relatively small, apart from building materials (wood products). In the United States, more than 90% of total organic chemical production is based on fossil feedstocks. Biomass accounts for less than 1% of all liquid fuels, essentially all of it ethanol derived from corn. Approximately 7% of the total U.S. corn crop is processed into fuel ethanol, industrial starches, industrial ethanol and other chemicals. Not withstanding the relatively small size of the biomass-derived chemicals and fuels industry, this industry produces a very wide range of products including oils, inks, pigments, dyes, adhesives, lubricants, surfactants, organic acids and many other compounds.

As we have seen, biomass is relatively low cost. However, the processes for converting biomass to industrial products are not, in general, well enough developed or low cost enough to compete effectively with comparable petroleum-derived products. Petroleum-derived products are supported by over a century of research, development and commercial experience. However, the competitive position of biomass is beginning to change. This change is being driven by a combination of technical, economic and social/political factors. Several recent authoritative reports suggest a gradual shift over this next century to a much larger fraction of fuels, chemicals and materials derived from biomass.

B. New and Developing Uses of Biomass

1. New Chemicals and Materials Uses

From 1983 to 1994, the sales of some biomass-derived products (fuel and industrial ethanol, corn syrups, citric acid, amino acids, enzymes and specialty chemicals, but excluding pharmaceutical products) rose from about $5.4 billion to approximately $11 billion. These products seem likely to continue to grow. The market for new and existing enzymes may be particularly strong, given the ability of enzymes to transform biomass into new products and to provide more environmentally clean chemistries for older petroleum-derived products and processes. Enzymes also have growing applications to improve environmental quality while reducing costs in selected agricultural and domestic uses (e.g., animal feed processing and detergents). Enhanced environmental compatibility and economic benefits are also key factors driving the adoption of soybean oil-based inks. These soybased inks were introduced in the 1970s in response to oil shortages and now account for about 40% of all inks.

In the United States, approximately 100 million tons per year of organic chemicals are produced annually, with much less than 10% of these chemicals currently derived from biomass. It seems likely that chemical uses of biomass will grow fastest among these or other organic chemicals, particularly for those chemicals that contain oxygen. Some examples of these oxygenated chemicals include organic acids and their derivatives (acetic, adipic, lactic and succinic acids and maleic anhydride), alcohols (butanol, isopropanol, propanediol, and butanediol) and ketones (acetone, methyl ethyl ketone). Indeed, the Cargill–Dow joint venture is focused on polymer production from lactic acid while the DuPont venture with Tate and Lyle is focused on 1, 3 propanediol as another polymer feedstock.

Therefore, bioplastics may prove to be the most rapidly growing new materials application for biomass. Industrial starches, fatty acids, and vegetable oils can serve as raw materials for bioplastics, including polymer composite materials. Waste paper and crop and forest wastes and virgin materials are being used as the basis of new composite materials and new fabrics, including Tencel, the first new textile fiber to be introduced in 30 years.

It is instructive to consider the amount of plant matter that might be consumed by new chemicals and materials uses of biomass. Given total U.S. production of about 100 million tons of organic chemicals annually, this is about one third of the total mass of the U.S. corn crop of approximately ten billion bushels per year. The corn residue or stover that might be converted to various products to substitute for or replace these organic chemicals is easily equal to the total mass of these organic chemicals, even without converting any of the corn itself. Furthermore, corn yields continue to increase.

If we assume a 1% per year increase in yield for corn (versus 3% per year over the past 50 years) and no change in the planted acreage, then the annual increase in corn produced is about 100 million bushels per year, or over 2 million metric tons of new corn every year. The Cargill–Dow Polymers plant being opened in Blair, Nebraska, in late 2001 will produce 140,000 metric tons per year of polylactic acid from approximately 200,000 metric tons of corn. That is, a new large scale plant for bioplastics will only use about 10% of one year's *increase* in the corn crop. Thus it seems unlikely that biomass use for chemicals and materials will really have much effect on grain supplies and prices. However, this does not hold true for new large scale liquid fuel uses of biomass.

2. New Liquid Fuels from Biomass

Total consumption of gasoline and diesel fuel in the United States is about 150 billion gallons per year. Assuming an average density of these liquid fuels of six pounds per gallon, the total mass of raw material that would need to be processed just to supply the U.S. with liquid fuel is about 450 million tons per year, assuming complete conversion of the mass of the raw material into fuel. In practice, significantly less than 100% raw material conversion to fuel products will be obtained.

In the case of biomass raw materials, however, the situation is even more constrained. If the entire domestic corn crop of about 10 billion bushels per year were converted to fuel ethanol, the total gallons of fuel produced would be less than 20% of domestic diesel and gasoline consumption. However, ethanol has an energy content about 70% of that of gasoline, a consequence of its oxygenated character. Thus a gallon of ethanol will provide lower vehicle mileage than a gallon of gasoline, even when burned in high compression engines designed to take advantage of its high octane value. Thus grain-derived ethanol can never meet more than a small fraction of our liquid fuel needs, however important corn ethanol may be as a bridge to a biomass fuel industry based on lignocellulosics. What is the potential of the lignocellulosics for liquid fuel production?

The United States produces at least 300 million tons per year of lignocellulosic crop and forest residues that might be available for conversion to liquid fuels. Current laboratory best yields of ethanol from such residues are about 100 gallons per ton. The total ethanol that might be produced from these lignocellulosic residues is therefore approximately 30 billion gallons per year, a significant fraction of our total domestic liquid fuel demand. Worldwide, if all of the corn stover, rice straw and sugarcane bagasse now produced were converted to ethanol at these yields, approximately 200 billion gallons of ethanol could be generated annually, versus an approximate worldwide consumption of petroleum-derived liquid transportation fuels of 800 billion gallons per year.

The requirements for liquid fuels obviously depend strongly on vehicle mileage. For example, if average vehicle fuel efficiency were to increase by a factor of two, demand for liquid fuels would be cut in half. Were that to happen, the fuel ethanol that could be produced just from the residues of these three major crops might satisfy nearly half of the worldwide demand for liquid transportation fuels, without the need to plant additional crops dedicated for energy uses. We must not lose sight of the need to work on the demand side of the fuel equation, as well as the supply side.

3. Land Requirements for Fuel Production

However, assuming no improvement in vehicle efficiency and that only half of the tonnage of the three major crop residues (corn stover, rice straw and sugarcane bagasse) is available for conversion to fuel ethanol (because of possible constraints of collection, erosion, alternative uses, etc), the approximate worldwide liquid fuel replacement that might be required from biomass-derived ethanol could be as high as one thousand billion gallons per year. At 100 gallons per ton, roughly 3.0 billion acres of land worldwide would need to be devoted to liquid fuel production from biomass to produce the total required biomass tonnage of 10 billion tons per year, assuming that the average U.S. hay production of 3 tons per acre per year were obtained.

However, as pointed out earlier, there is great potential to increase crop yields. Several tropical grasses, including sugarcane, have been shown to yield as much as 25–35 tons of dry matter per acre per year. Furthermore, most land currently used for animal pasture (about 7 billion acres worldwide) is not managed for intensive biomass production. For example, the United States has about 600 million acres in permanent pasture. Grazing animals obtain about 150 million tons of forage annually from this pasture, for an average yield of less than 0.3 tons per acre per year, versus about 3 tons per acre per year from our managed hay lands. Without a demand, there is no incentive to increase forage grass production on the available pasture land.

Therefore the demand for land and other agricultural resources required to support biobased industrial products is probably not a factor for chemicals and materials, but will be an issue for biobased fuels. The demand for land to supply liquid fuels depends on the yields of biomass from the land, the yield of fuel from the biomass and the miles traveled per unit of fuel. All three factors are important and must be considered in conjunction. Increasing the efficiency (yield) of each step will increase the overall system efficiency. In particular, biomass conversion to fuel ethanol must be highly efficient and low cost if the product is to compete with petroleum-derived fuels.

4. Cost of Liquid Fuel Production from Biomass

It is well known that fuel ethanol derived from corn must be subsidized to compete with gasoline. Raw material costs (primarily the corn) alone are in the neighborhood of $1.00 per gallon of ethanol produced, without any allowance for processing costs. Therefore, it seems unlikely that corn ethanol will ever be able to compete economically with petroleum-derived fuels. Nonetheless, a large

industry producing over 1.3 billion gallons of fuel ethanol from corn in highly integrated, efficient plants has arisen in the United States over the past twenty years, based at least partly on this subsidy.

It is not the purpose of this article to argue the pros and cons of subsidizing the corn ethanol industry. However, we note that the existence of the corn ethanol industry provides a learning opportunity, a production platform and a marketing springboard for both a chemicals from biomass industry as well as potentially huge industry based on converting lignocellulosic materials to fuel ethanol. Such an industry may eventually arise because lignocellulosic materials are so inexpensive.

While the cost and supply of grain severely limit its potential to replace a large percentage of gasoline, the situation for lignocellulose-derived ethanol is very different. Ample raw material supplies exist for lignocellulosic biomass. Further, because crop residues, grasses, hays and wood residues cost much less than grain, fuel ethanol produced from these lignocellulosic materials can potentially cost much less than grain ethanol. Assuming biomass processing technology at a similar stage of maturity as petroleum processing technology, it has been shown that the cost of fuel ethanol from lignocellulosics should be in the \$0.50–\$0.70 per gallon range, assuming biomass costing about \$30.00 per ton delivered. Given low cost corn stover, another estimate projects ethanol costs at less than \$0.50 per gallon for large scale plants, assuming best laboratory yields. Clearly, there is real potential for low cost fuel ethanol from cellulosic biomass *if the processing technology for conversion of biomass can be made both efficient and inexpensive.* Processing technology is the subject of the final section of this article.

IV. BIOPROCESSING OF BIOMASS

A. Historical Lessons from the Chemical and Petroleum Processing Industries

1. Importance of Raw Material and Processing Costs for Commodities

The chemical and petroleum industries grew together over the past century. These industries add value to raw materials by converting them to commodity and speciality products. Processing technologies of various kinds are used including chemical, thermal, physical and biological methods. By long experience, it has been found that the cost to produce commodities depends on two major factors: (1) the cost of the raw material and (2) the cost of the conversion process. The industries that produce chemicals and fuels from petroleum are characterized by high raw material costs relative to processing costs. Typically 50–70% of the cost to produce a commodity product from petroleum is due to the petroleum cost itself. This is why gasoline prices fluctuate so widely when crude oil prices change.

However, for the analogous biobased products industries, the processing costs predominate, rather than raw material costs. Therefore, a given percentage decrease in processing costs has much more impact on the profitability and economic competitiveness of biobased industrial products than does the same percentage decrease in raw material costs. As we have seen, the cost per ton of biomass raw materials is generally comparable to (e.g., corn grain) or much less (e.g., corn stover) than the cost of petroleum. Because of this fact, there is real potential for biobased products to be cost competitive with petroleum products if we can learn how to reduce the costs of processing biomass to desired products. Before discussing specific technical areas that seem to offer the best opportunities to reduce processing costs, a brief discussion of several lessons from the petroleum and chemical industries will be useful.

2. Need for Complete Raw Material Utilization

This point is so elementary that it is often overlooked. For processes producing millions of pounds of biobased plastics or billions of gallons of fuel ethanol per year, essentially all of the raw material must be converted to saleable products, or at a minimum, not into wastes requiring expensive treatment and disposal. The petroleum refining industry has over time learned how to convert nearly all of the raw material into products. To compete effectively with this entrenched industry, the biobased products industry must become similarly efficient. Yield (conversion of raw material to products) must be increased and improved.

A simple calculation will illustrate this point. If a raw material costing \$0.10 per pound is converted into product at a yield of 0.9 pounds of product per pound of raw material, then the raw material cost is about \$0.11 per pound of product. If the same raw material is converted to product at a yield of only 0.5 pounds of product per pound of raw material, the raw material cost is now \$0.20 per pound of product, nearly double the previous case. The petroleum industry is characterized by high yields; the biobased products industry must strive to improve its yields also.

While low yields are a definite economic handicap, they may be an even more severe environmental (and by consequence an economic) handicap. Whatever portion of the raw material is not converted to saleable products becomes waste instead. These wastes must be treated before disposal, if disposal is possible. Liquid wastes from biobased products will likely be characterized by relatively low

toxicity, but by high oxygen demands if processed in conventional sewage treatment facilities. Solid wastes from biomass processing could occupy large volumes of landfill space, and would tend also produce high oxygen demand liquid effluents from landfills.

The volume of landfill space that might be occupied by these products is remarkable, particularly for biobased fuels. For example, corn stover contains approximately 10–12% of protein and fat, in addition to the 70–75% of the plant material that is carbohydrate and that could be converted to fuel ethanol in a large fermentation facility. If no economic uses are found for the protein and fat or if economical recovery processes are not developed, these biomass components will have to be burned, landfilled, or treated as sewage and the resulting sludge disposed of. A hypothetical ethanol plant producing 100 million gallons of ethanol per year (less than 0.1% of current U.S. gasoline demand) from 1 million tons of corn stover will also produce at least 100,000 tons of protein and fats. Assuming a bulk density of about 50 lb per cubic foot, this much protein and fat will occupy a volume equivalent to the surface area of a football field stacked approximately 100 ft deep. Clearly there is strong economic incentive to find uses for all of the components of biomass.

Therefore, as has occurred with the oil refining industry, "biorefineries" will tend to emerge. These biorefineries will be large, highly integrated processing plants that will yield numerous products and will attempt to sell a pound of product for every pound of raw material entering the plant. Prototype biorefineries already exist, including corn-wet and dry mills, soybean processing facilities and pulp and paper mills.

The number and variety of these biobased products will increase over time, as has occurred with the oil refining industry. Many biorefinery products can also be produced by oil refineries; including liquid fuels, organic chemicals and materials. However, biorefineries can make many other products that oil refineries cannot, including foods, feeds, and biochemicals. These additional capabilities give biorefineries a potential competitive edge and may provide increased financial stability.

3. Incremental Process Improvement

As we have seen, the number of products from a refinery tends to increase with time. In addition, the processing technologies used by refineries tend to improve incrementally over time. This is partly due to research that improves or replaces existing processes, supported within the cost structure of a successful industry. Research targets those process elements that most impact the overall cost of conversion. Incremental cost reduction is also partly due to organizational learning that can only occur in a functioning industry. The more biorefineries that are established and become successful, the more that will be built as risks decline and "know how" increases.

The cumulative effect of incremental process improvement is to cause the raw material costs to eventually become the dominant cost factor. This has already occurred with the oil refining industry and will take place in the biomass processing industries as these are established and grow to maturity. In this regard, biorefineries have a significant potential advantage over petroleum refineries because plant-based raw materials are abundant, widely available and inexpensive. The availability and prices of plant raw materials may thus be more stable and predictable than those of petroleum. As we have seen, plant raw material prices are already comparable on a cost per ton basis with petroleum and coal. Over time, petroleum prices must rise to reflect the fact that it is a nonrenewable resource, while there is the potential to keep biomass costs low indefinitely.

4. Innovation and Risk

While biorefineries have some inherent advantages over petroleum refineries, they also have some significant disadvantages. First, the costs and risks of petroleum refining are well understood and the commodity organic chemicals industry based on petroleum is very well developed. However, the costs and risks of biomass refining are not nearly so well understood, particularly for large scale plants converting lignocellulosic biomass into fuel ethanol. Innovation is always regarded as risky compared to the *status quo*. Investors demand a greater return on investment to compensate for these increased risks and they also require additional capital investment in the processing plant to reduce processing risks or uncertainties. Second, when the petroleum industry was new there was little or no competition for many of its products as they emerged. However, most of the potential new biobased products must compete with a comparable petroleum-derived product. These two factors are significant hurdles for biomass processing industries to overcome.

A better fundamental understanding of the underlying science and technology for biomass conversion processes could reduce the risks of such processes as they are scaled up to commercial practice. Better fundamental understanding would reassure investors and allow them to reduce the return on investment required. Better fundamental understanding of the processes would also tend to reduce the processing equipment redundancy required because of lack of certainty about its proper functioning under commercial practice. Some of the key areas of biomass processing in which greater fundamental understanding is required are discussed below.

B. Current Status of Biomass Processing

Biomass processing to industrial products based on starch, sugar, and oilseed raw materials is partially developed. Fiber crop processing to pulp and paper is very well developed and is not discussed further here. Processing of lignocellulosic materials to industrial products other than pulp and paper is very limited.

For processing of starch, sugar, and oilseed crops, a primary need is to develop additional industrial products from these raw materials. This is because a processing infrastructure, or at least the beginnings of one, already exists or could be developed for each of these raw materials. Therefore the capital risk of a totally new plant is not required. Corn wet and dry mills, sugar refineries for cane and beet sugar, and oilseed crushing mills already exist and industrial products are already produced from these raw materials.

As additional products are developed and appear profitable, they can often be added to existing product lines at existing facilities, for instance in a corn wet mill. Corn wet millers have steadily increased the number and variety of their products over the past two decades or so. This trend is likely to continue. Growth of new biobased products at oilseed crushing mills appears to have been much slower.

However, when circumstances appear favorable, totally new processing facilities can and will be built around specific new industrial products, as evidenced by the new plants for polylactic acid production announced and/or under construction. Many of these starch or sugar-based products might also be produced at even lower cost from inexpensive sugars generated in a lignocellulose conversion plant. Processing plants based on lignocellulose are struggling to become established, at least partly because the processing technology is underdeveloped and therefore relatively expensive compared to petroleum processing technology.

C. Priorities for Developing Lignocellulose Biorefineries

1. Enhancing Yield and Feedstock Modification

The emphasis in this article on development of processing technologies for large scale refining of lignocellulosic materials to industrial products is not intended to detract from at least two other important development areas that are not directly connected with processing technology. The first of these is yield. The profitability of biomass conversion industries using dedicated (grown for that purpose) feedstocks will be strongly affected by the yield of raw material (tons of dry lignocellulosic material produced per year per acre of land). Agronomic research to increase yields and develop improved plant varieties for biomass production is important and requires continuing attention. A key area of agronomic research, particularly for large-scale energy crops, is to reduce the amounts of fertilizers, pesticides, herbicides and other inputs required, both to minimize costs and to reduce potential environmental hazards.

The second important area that is not directly process development is plant feedstock modification, either by breeding or genetic modification. As mentioned, feedstocks can be altered to make them easier to process, a major advantage of biomass feedstocks compared with petroleum raw materials. Feedstocks can also be altered to contain larger amounts of desirable components or even to contain entirely new components such as bioplastics. Careful integration of processing technology development and product recovery with feedstock modification is required to achieve the maximum benefits of these genetic manipulations. Presumably yields might also be affected by some of these modifications. Thus feedstock modification and yield enhancement should proceed as an integrated whole.

2. New Technologies Needed for Low Cost Lignocellulose Conversion

While it is relatively easy to convert starchy materials such as corn to fermentable sugars and then to a variety of fermentation products, lignocellulosic materials are much more difficult to convert to fermentable sugars. The potential benefit of economical conversion of lignocellulosics to fermentable sugars is the much larger possible volumes and therefore lower costs of lignocellulose-derived sugars compared to starch-derived sugars. Such low cost sugars are a prerequisite to inexpensive fuel ethanol. Inexpensive sugars could also significantly reduce the costs of other biobased products such as polylactic acid or 1,3 propanediol.

Three primary areas for new technology development are required to reduce the cost of producing fuel ethanol and commodity chemicals from lignocellulosic materials: (1) an effective and economical pretreatment to unlock the potentially fermentable sugars in lignocellulosic biomass or alternative processes that enable more biomass carbon to be converted to ethanol or other desired products, (2) inexpensive enzymes (called "cellulases") to convert the sugar polymers in lignocellulose to fermentable sugars, and (3) microbes that can rapidly and completely convert the variety of five and six carbon sugars in lignocellulose to ethanol and other oxygenated chemicals.

Several lignocellulose pretreatment processes have recently been developed that promise to be technically effective and affordable, and some of them are undergoing large scale testing and development. Advanced lignocellulose treatments include processing with ammonia and

other bases, pressurized hot water and catalytic amounts of acid. Such pretreatments may eventually make it possible to convert a large array of lignocellulose residues into useful products. It has long been recognized that a technologically and economically successful lignocellulose pretreatment will not only unlock the sugars in plant material for conversion to industrial products but will also make these sugars more available for animal feeding.

Considerable progress has been made in developing genetically engineered microorganisms that can utilize the complete range of sugars in lingocellulosic materials. Both genetically engineered bacteria and yeasts are now available that utilize both five and six carbon sugars and convert them to ethanol. However, less progress is apparent in production of low cost cellulase enzymes, an active area of development at this time. Fortunately, while research on cellulases will take both time and money, the research pathways required to achieve success, (i.e., much lower cost active cellulases), are relatively clear.

It may also be possible to largely bypass the processes of cellulose pretreatment and enzymatic hydrolysis to produce fermentable sugars. One such approach is to gasify biomass to mixtures of carbon dioxide, hydrogen and carbon monoxide and then to ferment the resulting gas mixture to ethanol or other products. This amounts to a technological "end run" around the processes of biomass pretreatment and enzymatic hydrolysis, and could also significantly simplify the fermentation process. Two major technical obstacles to this approach include clean up of the gas stream, particularly to remove sulfur and nitrogen-containing compounds, and also the difficulty of transferring slightly soluble gases into a liquid fermentation mixture.

3. Generic Biomass Processing Technologies

The following comments apply generally to biomass conversion, not just to lignocellulose conversion. Processing technologies that utilize microbes and enzymes have great potential for low cost biomass processing. Unlike most thermal and chemical processes, bioprocesses take place under relatively mild conditions of temperature and pressure. Higher temperatures and pressures add significantly to the cost of processing in conventional chemical industries so that advanced bioprocessing technologies have the potential to be less expensive than their non biological counterparts. Some advanced bioprocessing technologies utilizing microbes and enzymes have already been developed, for example, immobilized cell technology and simultaneous hydrolysis and fermentation of sugars from lignocellulosics. Bioprocesses result in stereospecific conversions (the particular arrangement of atoms in space) and produce relatively nontoxic byproducts. However, the volume of such byproducts can be very large and there is a pressing need to find markets for all such products and byproducts.

Bioprocessing research should therefore focus on (1) increasing processing rates to reduce the capital investment required, (2) increasing product yields to decrease the raw material costs and to reduce the load on the separation and waste disposal systems, and (3) increasing product concentrations to reduce separation costs. One drawback is that bioprocesses typically yield dilute aqueous product streams that require further processing to separate and purify products. Separation technologies for biobased products are typically less developed than separation technologies for comparable petroleum-based products. A major need is to find low cost ways of removing the large amounts of water that often accompany biobased products. In general, research on the underlying production processes should focus on the science and engineering required to overcome the most significant cost barriers to commercializing biobased products.

Experience with commercial amino acid production illustrates the potential of combining inexpensive raw materials with advanced processing technologies. International amino acid markets were completely dominated by Japanese firms in the early 1980s. However, in the 1990s U.S. companies used inexpensive corn-based sugars and an advanced processing method, immobilized cell technology, to penetrate these markets and now occupy a significant part of the global amino acid trade.

One of the reasons these corn-based sugars for amino acid production were so inexpensive is because they were produced in large, integrated biorefineries. The Archer Daniels Midland plant in Decatur, Illinois, is a prototypical biorefinery. At that location, a large corn wet-milling plant and a steam and electricity cogeneration station burning waste tires form the nucleus for several other plants that are highly integrated. These other plants are an ethanol facility as well as an amino acid production plant. Biorefineries, whether based on corn, straw or any other material, must aspire to a similar degree of integration and effectiveness in raw material conversion.

V. POTENTIAL AND LIMITATIONS OF BIOMASS AND BIOBASED INDUSTRIAL PRODUCTS

A. Potential Benefits

Biomass production and processing have the potential to give us a uniquely sustainable source of organic chemicals, organic materials (such as biopolymers) and liquid transportation fuels. Biomass can also help us sustainably

produce electricity, although there are several other sustainable sources of electricity. Biomass is also uniquely a source of human food and animal feed.

Because biomass production is widely dispersed geographically, biobased industrial products can potentially form the basis of both local and worldwide economic systems that are much more equitable and balanced. Also because biomass production is widely dispersed, resource-driven international conflicts over petroleum might be minimized or avoided entirely.

Biomass production is a key part of global cycles of carbon, oxygen, nitrogen, water and other compounds. If we intelligently produce and use biobased industrial products we may actually improve environmental quality and increase or enhance the stocks of "natural capital" such as soil, water and air upon which all life depends. Numerous opportunities also exist to integrate biomass production and processing with waste utilization and recovery of damaged or less fertile lands. For example, the organic fraction of municipal solid wastes might be combined with human and animal wastes and composted to enrich marginal soils producing perennial grasses for a bioethanol facility. Furthermore, since plants fix atmospheric carbon both in their above and below ground parts, the potential exists to continue to use petroleum and other fossil fuels indefinitely, balancing the amount of atmospheric carbon dioxide liberated by fossil fuel combustion with the uptake and fixation of carbon dioxide by plants.

B. Potential Limitations of Biomass and Biobased Industrial Products

Perhaps the most serious potential limitation of biomass and biobased industrial products is the possible conflict with food and feed production on the same land. While biomass utilization for organic chemicals and materials, done properly, is not likely to result in conflicts with food production, biomass production and utilization for liquid fuels such as ethanol might indeed conflict with food production. This is particularly true if fuel use efficiency does not increase dramatically over the time frame that biofuels are implemented. Food production will always be a higher human priority than production of fuels or plastics. This issue must be carefully considered and appropriate resolutions achieved if biobased industrial products, including biomass-derived liquid transportation fuels, are to provide us their full social, economic and environmental benefits.

Some threats to biodiversity and water and soil quality are also possible from greatly increased levels of biobased industrial products. Erosion and increased contamination of soil and water with fertilizers, pesticides and herbicides might result from intensive production of biomass for biobased products. Disposal of biobased products, done with reasonable care, should not have environmental impacts more severe than the corresponding petroleum-derived products. In fact, biobased products are well suited to composting or other resource recovery approaches that return their constituent atoms to the global cycles of materials.

C. Achieving the Benefits of Biobased Products

While the potential benefits of biobased products are certainly real, so are their limitations and possible problems. One way of achieving the benefits of biomass processing to biobased products is to do careful, system-level studies of specific products in order to anticipate and resolve potential problems before large industries are launched and the damage is done. Life cycle analysis is suited to such system studies. For example, there is an obvious potential for biomass production for biobased products to conflict with food production. Careful studies are required to anticipate and resolve such conflicts before they occur.

One potential resolution of this apparent conflict with food and fuel production is to coproduce foods and animal feeds with fuel and chemical production from biomass. Most biomass produced is actually fed to animals, rather than directly to humans. Since most biomass also contains protein (required in all animal diets), the potential exists to recover this biomass protein in a biorefinery and use it to feed animals, or perhaps even people. Assuming an average protein content of 10% in grasses and residues, and assuming 80% recovery of this protein in biorefineries also producing ethanol fuel, about 1 billion tons of grass would be required to replace all of the protein currently produced worldwide as high protein meals from oilseeds such as soybeans. The equivalent amount of ethanol produced would be about 100 billion gallons per year, about half of the U.S. demand for liquid transportation fuels.

Similarly, the calories (food energy) in lignocellulosic materials are not very available for animal digestion and they are essentially useless in human nutrition. However, if the technical roadblock of lignocellulose pretreatment for production of fuels is resolved, it will also be resolved for pretreatment to increase the food and feed calories available from lignocellulosics. For example, ruminant animals typically digest less than half of the calories potentially available in grasses. If pretreatments make those calories 90% percent available both for fermentation to ethanol and also for animal feeding, then the treatment of about 4 billion tons per year of grasses will make available for feed (or food) and fuel uses new, additional calories approximately equal to the calories contained in the entire world grain crop of about two billion tons per year. Thus while both

the protein and calorie issues in biomass processing need careful analysis, it may be possible to actually increase, rather than decrease, world food and feed resources as a consequence of large-scale biomass processing to fuels.

While the potential difficulties of biomass processing are certainly real, so are the potential benefits. In essence, we have the unique opportunity to guide and develop a new industry so that it improves the quality of our environment at the same time it provides many economic, environmental and social benefits. Both the opportunity and the challenge are ours.

SEE ALSO THE FOLLOWING ARTICLES

Bioenergetics • Biomass Utilization, Limits of • Bioreactors • Energy Efficiency Comparisons Between Countries • Energy Flows in Ecology and in the Economy • Energy Resources and Reserves • Waste-to-Energy Systems

BIBLIOGRAPHY

Hawken, P., Lovins, A., and Hunter Lovins, L (1999). "Natural Capitalism," Little, Brown and Company, Boston.

Klass, D. L. (1995). "Fuels from Biomass," *In* "Encyclopedia of Energy Technology and the Environment," (A. Bisio and S. Boots, eds.), Vol. 2, pp. 1445–1496, Wiley, New York.

Lugar, R. G., and Woolsey, R. J. (1999). "The new petroleum,"*Foreign Affairs* **78**(1), 88–102.

Lynd, L. R. (1996). "Overview and evaluation of fuel ethanol from cellulosic biomass: Technology, economics, the environment and policy," *Annu. Rev. Energy Environment* **21,** 403–465.

Lynd, L. R., Wyman, C. A., and Gerngross, T. U. (1999). "Biocommodity engineering," *Biotechnology Progr.* **15**(5), 777–793.

McLaren, J., and Faulkner, D. (1999). "The Technology Roadmap for Plant/Crop-Based Renewable Resources 2020," U.S. Department of Energy, Washington, DC, DOE/GO-10099-706.

Morris, D., and Ahmed, I. (1992). "The Carbohydrate Economy: Making Chemicals and Industrial Materials from Plant Matter," Institute for Local Self-Reliance, Washington, DC.

National Research Council (2000). "Biobased Industrial Products: Priorities for Research and Commercialization," National Academy Press, Washington, DC.

Biomaterials, Synthetic Synthesis, Fabrication, and Applications

Carole C. Perry
The Nottingham Trent University

I. Introduction to Medical Biomaterials
II. Aspects of the Structural Chemistry of Natural Materials Used in the Human Body
III. General Repair Mechanisms and Biocompatibility
IV. Materials of Construction
V. The Way Forward, Tissue Engineering

GLOSSARY

Biomaterials A substance that is used in prostheses or in medical devices designed for contact with the living body for an intended application and for a planned time period.

Biocompatibility The capability of being harmonious with biological life without causing toxic or injurious effect.

Ceramics Stiff, brittle materials that are generally prepared by high temperature methods; the resulting materials are insoluble in water.

Composites Materials composed of a mixture or combination of two or more microconstituents or macroconstituents that differ in form and chemical composition and that are essentially insoluble in each other.

Metals Any of a class of elements that generally are solid at ordinary temperature, have a grayish color and a shiny surface, and will conduct heat and electricity well.

Polymers Large molecules formed by the union of at least five identical monomers.

Tissue engineering The study of cellular responses to materials implants, manipulation of the healing environment to control the structure of regenerated tissue, the production of cells and tissues for transplantation into the body, and the development of a quantitative understanding of biological energetics.

I. INTRODUCTION TO MEDICAL BIOMATERIALS

The normal description of a biomaterial is "a substance that is used in prostheses or in medical devices designed for contact with the living body for an intended application and for a planned time period." The development of

biomaterials has occurred in response to the growing number of patients afflicted with traumatic and nontraumatic conditions. As the population grows older there is an increased need for medical devices to replace damaged or worn tissues. The market is a billion dollar per year market and requires the skills of clinicians, surgeons, engineers, chemists, physicists and materials scientists to work cooperatively in the development of materials for clinical use.

There are five classes of biomaterials; metals, ceramics, biological materials/polymers, synthetic polymers, and composites. The choice of material to replace biological tissue is largely governed by the physical properties of both the natural tissue and the proposed replacement. In general, natural and synthetic polymers are used to replace skin, tendon, ligament, breast, eye, vascular systems, and facial tissues, and metals, ceramics and composites are used to replace or reinforce bone and dentin. Replacements for these natural tissues clearly require materials of different strength. Table I shows the strength of the main groups of natural materials together with the synthetic counterparts used in the development of replacement materials. It is often the case that the strengths of the materials used to replace natural components are stronger and/or stiffer which often leads to problems of compatibility both in respect of mechanical behaviour of the implant within the host and in terms of the biologic response.

TABLE I Physical Properties of Tissues and Materials Used in Their Replacement

Material	Ultimate strength (Mpa)	Modulus (MPa)
Natural materials		
Soft tissue		
Arterial wall	0.5–1.72	1.0
Hyaline cartilage	1.3–1.8	0.4–19
Skin	2.5–16	6–40
Tendon/ligament	30–300	65–2500
Hard tissue (bone)		
Cortical	30–211	16–20 (GPa)
Cancellous (porous)	51–193	4.6–15 (GPa)
Synthetic materials		
Polymers		
Synthetic rubber	10–12	4
Glassy	25–100	1.6–2.6 (GPa)
Crystalline	22–40	(0.015–1) (GPa)
Metal alloys		
Steel	480–655	193 (Gpa)
Cobalt	655–1400	195 (GPa)
Platinum	152–485	147 (GPa)
Titanium	550–680	100–105 (GPa)
Ceramics		
Oxides	90–380 (Gpa)	160–4000 (GPa)
Hydroxylapatite	600	19 (GPa)
Composites		
Fibers	0.9–4.5 (Gpa)	62–577 (GPa)
Matrices	41–106	0.3–3.1

Many of the materials used today in the clinical environment were not originally engineered for biomaterials applications. In crude terms, they became “biomaterials” when, by a series of trial-and-error experimentation, they were implanted in the human body in a variety of forms and found to “work.” Clearly there were many other materials which did not ‘work’ causing at the least perhaps pain and discomfort to patients and at the worse unnecessary suffering and death. Increase in litigation for problems allegedly caused by biomaterials has caused some companies to remove products from sale in this area and may lead to a distinct shortage of available materials for device fabrication. For a variety of reasons, trial-and-error optimization is not the way forward for the production of the next generation of materials to be used in the human body.

What is required is the systematic design of a wide range of materials with specific functions for predefined medical use. The goal is to produce biomaterials that smoothly integrate into living systems rather than fighting against the normal functioning of a living body. There are many factors to consider in order to understand how this might be accomplished. The factors include the structure of the original material to be replaced or augmented, its physiology, anatomy, biochemistry and biomechanics including pathophysiological changes that have necessitated the use of a substitute biomaterial. In addition, as devices are often present in the body for considerable periods of time then it is necessary to understand the natural degenerative processes of normal tissues, particularly in relation to the biomaterial substitute. This latter area is at present very poorly understood. All of the above clearly impact on the design and development of materials for clinical usage. Thus materials need to be developed with a *clear* understanding of the nature and extent of interactions between the device (whatever it is) and the surrounding tissue. It cannot be emphasised too strongly the importance of biocompatibility in the development of the next generation of materials for applications in a biological environment.

This chapter will describe the materials currently used as biomaterials and routes to their formation. It will describe some aspects of the structural chemistry of natural materials that are to be replaced or augmented and it will look at the way forward for the design of materials for use in the medical environment in the 21st century.

II. ASPECTS OF THE STRUCTURAL CHEMISTRY OF NATURAL MATERIALS USED IN THE HUMAN BODY

Biological organisms make use of proteins, polysaccharides and combinations of these two types of molecule in the polymeric phases that are found in a living organism together with simple calcium salts. Chemical composition

and structure both play important roles in determining the properties of the natural polymeric phase whether it is used alone or in a composite form.

A. Natural Polymers

The key to the mechanical properties of fibrous materials (collagen, silks, chitin and cellulose are all natural examples and polyethylene and nylon are examples of synthetic polymers) lies in their structure, in particular the extent of their crystallinity. Crystallinity equates with material rigidity, hardness and toughness, Table I. For crystalline polymers to form they must have great regularity of chain structure and be linear in form. The presence of more than one stereo-isomeric form of the polymer can cause problems in structural alignment for synthetic polymers but this is not a problem in natural polymers. Collagen or cellulose as only one stereo-isomer is utilized in the building of these structures. When the level of crystallinity is less than 20%, the crystalline regions act as cross-links in an amorphous polymer network with materials being mechanically similar to cross-linked rubbers. Above 40% crystallinity the materials become quite rigid and the mechanical properties of the polymers are largely time-independent. The behavior of drawn fibers, such as those produced from synthetic polymers may be further different from the bulk material as the regions of amorphous character are greatly extended and aligned perpendicular to the stress direction. In general, the Young's modulus (the stress–strain curve) of a bulk crystallized polymer is two to three orders of magnitude greater than that of an amorphous polymer such as rubber and the modulus of oriented, crystalline fibers is another order of magnitude greater still.

Collagen is a prime example of an important crystalline polymer in mammals and is found in both rigid (e.g., bone) and pliant materials (e.g., tendon, skin and cartilage) (Table II). Each collagen chain is ca. 1000 amino acids in length and is dominated by a 338 contiguous repeating triplet sequence in which every third amino acid is glycine. Proline and hydroxyproline together account for about 20% of the total amino acids. The structure of collagen is based on the helical arrangement of three non-coaxial, helical polypeptides, stabilized by inter-chain and intra-chain hydrogen bonds with all the glycines facing the center of the triple helix. Individual collagen molecules are ca. 280 nm in length and 1.5 nm in width with adjacent molecules being transposed in position relative to one another by one quarter of their length in the axial direction. Collagen is the protein used in tissues where strength or toughness is required. In addition to the effect of its intrinsic chemical structure on the observed mechanical properties, part of its toughness arises from specific cross-links between the molecules. These naturally increase with age leading to a less flexible material as someone gets older.

Elastin is used in situations where a highly elastic fiber is needed such as in ligaments and blood vessels (Table II). The polypeptide chain of elastin contains high levels of both glycine and alanine, is flexible and easily extended. To prevent infinite extension the peptide sequence contains frequent lysine side chains that can be involved in cross-links that allow the fibers to "snap-back" on removal of tension. Elastin is only ever found as fine fibers and is usually found in association with collagen and glycosaminoglycans. The hydrophobic character of the protein (95% of the amino acids have nonpolar side chains) leads to water becoming more ordered and the polymer less ordered when fibers of the material are stretched.

Carbohydrate-based natural polymers include *glycosaminoglycans* (polysaccharides containing amino acids formerly known as mucopolysaccharides), *proteoglycans* (covalently bonded complexes of protein and glycosaminoglycan in which the polysaccharide is the dominant feature) and *protein–polysaccharide complexes* where a complex is held together through noncovalent linkages. Important examples are the *chondroitin sulfates* and *keratan sulfates* of connective tissue, the *dermatan sulfates* of skin and *hyaluronic* acid. All are polymers of repeating disaccharide units in which one of the sugars is either N-acetylglucosamine or N-acetylgalactosamine. Examples are given in Table II. A major function of this group of molecules is the formation of a matrix to support and surround the protein components of skin and connective tissue. A schematic view of the proteoglycan complex in cartilage used to bind to collagen in a strong network is shown in Fig. 1. The filamentous structure contains at its heart a single molecule of hyaluronic acid to which is attached via noncovalent interactions proteins known as "core" proteins. These in turn have chondroitin sulfate and

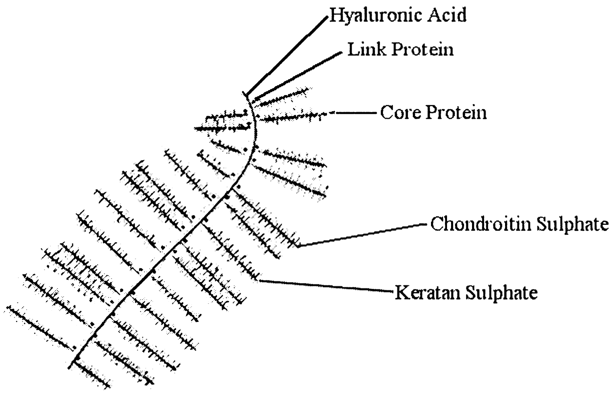
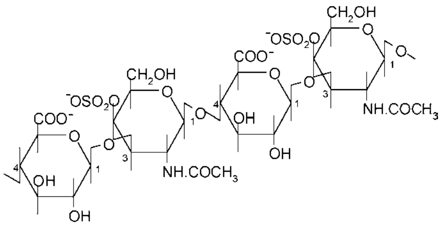
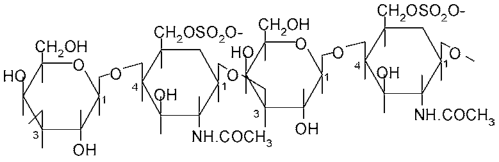
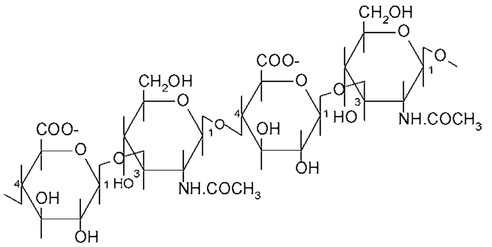


FIGURE 1 A schematic view of the proteoglycan complex in cartilage used to bind to collagen in a strong network.

TABLE II

Natural polymer	Structural unit
Collagen	(gly-x-pro/Hpro)$_n$ where x is another amino acid.
	Triple helices are built up from the basic structural units and held together by inter- and intra-molecular bonds between helices to form fibers
Elastin	(gly-val-gly-val-pro)$_n$
	Polypeptide rich in glycine and alanine, 95% hydrophobic residues
	Lysine side-chains used in cross-linking to give fibers
Chondroitin sulphate	Polymer of glucuronic acid and sulfated n-acetylglucosamine
	
Keratan sulphate	Polymer of galactose and sulfated n-acetylgalactosamine
	
Hyaluronic acid	Polymer of glucuronic acid and n-acetylglucosamine
	

keratan sulfate chains covalently bound to them. The properties of these molecules and their protein–polysaccharide complexes are largely determined by the poly-anionic character of the glycosaminoglycans. Both carboxyl and sulfated groups are ionized at physiological pH to give highly charged polymeric molecules where the molecule takes on a highly expanded form in solution. Perhaps unusally, glycosaminoglycan complexes of connective tissue may also contain a small amount of silicon (ca. 1 silicon atom per 100 sugar residues) as a cross-linking agent between adjacent chains.

Hyaluronic acid has additional functions within the body due to its high solubility in water. In addition to the properties described above, molecules are able to interact with one another at low concentrations to form entanglement networks and produce solutions with viscoelastic properties. In addition, if some permanent cross-bridges can form then gel-like structures with rubber-elastic properties can result. Hyaluronic acid is thus able to act as a viscosity modifier and a lubricating agent in the synovial fluid of joints and in the vitreous humor of the eye.

B. Natural Pliant Composites

Pliant composites are tendon, skin, and cartilage, all of which contain fibers (in some proportion) and a chemical matrix to support and modify the behavior of the high strength material. Collagen is generally used as the fibrillar component with differences in the thickness of these fibers (15 to 150 nm) being related to the required mechanical properties of the composite. There are differences in the extent of cross-linking between the collagen molecules and in the nature and organization of the fibrils and the matrix in which it is found.

Tendon is the structure that enables the rigid attachment of muscle to bone, and as such it must transmit the muscle

force with a minimum of loss. This is achieved through the parallel arrangement of collagen fibers to form rope-like structures with a high modulus of elasticity and high tensile strength. Tendons, such as the Achilles tendon, that are under a lot of stress probably contain collagen fibers that are more highly cross-linked to reduce the rate of stress–relaxation to an insignificant level.

Skin is a complex tissue made up of a thick collagenous layer (the dermis), a basement membrane and an overlying keratinized epidermal layer. The mechanical properties of skin arise principally from the dermis which is a three-dimensional feltwork of continuous collagen fibers embedded in a protein–polysaccharide matrix rich in both dermatan sulfate and hyaluronic acid, the latter being used to reduce frictional wear between the collagen fibers. Elastin fibers are distributed throughout the tissue or concentrated in the lower layers of the dermis depending on the precise location of the skin within the body. The arrangement of elastin fibers within a collagen framework results in a material showing rubber-elastic properties at small extensions but is limited at longer extensions by the dimensions of the collagen framework.

Cartilage acts as a material that maintains the shape of ears, the nose, and the invertebral disc. Cartilage contains collagen fibers, a proteoglycan matrix phase rich in chondroitin 4- and 6-sulfate and sometimes elastin fibers and yet the material must be able to resist compression and bending forces. Cartilage can be thought of as a hydrostatic system in which the fluid element is provided by the hydration water of the proteoglycan gel and the container provided by the collagen fiber meshwork which immobilizes the molecules of this gel. Thus, the rigidity of the system arises from the osmotic swelling of the proteoglycan gel against the constraints imposed by the collagen fiber system. Cartilage may additionally be mineralized and will be discussed below in conjunction with other mineralized tissues.

C. Natural Mineralized Tissues, Bone, Cartilage, and Enamel

Vertebrates construct their skeletal and dental hard parts from calcium phosphates with calcium carbonates being used for balance organs and egg shells. Bone, dentin, enamel, and mineralized cartilage all contain crystalline calcium apatite phases but the crystals exhibit different sizes, compositions, and ultrastructural organization. Apart from enamel they all contain collagen fibers, and additional inorganic salts and biomolecules.

Bone has unusual physical and mechanical properties in that it is able to support its own weight, withstand acute forces, bend without shattering and can be flexible without breaking within predefined limits. Bone also acts as an ion reservoir for both cations and anions. In material terms bone is a three-phase material; the organic fibers (collagen) can be compared to steel cables in reinforced cement, the inorganic crystalline phase (carbonated hydroxyapatite) to a heat-treated ceramic and the bone matrix to a base substance which performs various cellular functions. The unique physical and mechanical properties of bone are a direct result of the atomic and molecular interactions intrinsic to this unusual composite material.

TABLE III Polymers Used in Medical Devices

Polymer	Medical device applications
Polyethylene	Hip, tendon/ligament implants and facial implants
Polyethylene terphthalate	Aortic, tendon/ligament and facial implants
Polymethylmethacrylate	Intraocular lens, contact lenses and bone cement
Polydimethylsiloxane	Breast, facial and tendon implants
Polyurethane	Breast, vascular and skin implants

Bone comprises collagen fibrils intimately associated in an orderly fashion with small calcium phosphate crystals. The crystals are of carbonated-hydroxyapatite more correctly described as the mineral dahllite. The formula $Ca_{8.3}(PO_4)_{4.3}(CO_3)_x(HPO_4)_y(OH)_{0.3}$ represents bone mineral with the values of X and Y changing with age (Y decreases and X increases with increasing age, whereas X+Y remains constant with age equal to 1.7!). Traces of other elements such as silicon may also be associated with deposition of the mineral phase. The individual crystals have an average length of 50 nm (range 20–150 nm), width 25 nm (range 10–80 nm) and thickness of 2–5 nm. In addition to collagen at 85–90% of the detectable protein there are more than 200 noncollagenous proteins (NCPs) present. The three major classes of NCP's are acidic glycoproteins, proteoglycans and Gla- (γ-carboxyglutamic acid) proteins. The acidic glycoproteins contain considerable amounts of the amino acids phosphoserine, phosphothreonine, and γ-carboxyglutamic acid. The phosphopoteins are intimately associated with the initiation and regulation of crystal growth and may serve as a source of inorganic phosphate on enzymatic release by phosphatases. The proteoglycans have one or more (negatively) charged glycosaminoglycan chains attached to the main protein chain and may be present to inhibit crystal growth due to their negative charge and to reserve the extracellular space for future calcium phosphate crystal growth due to their ability to structure water. Both these classes of proteins together with alkaline phosphatase are found in a range of mineralized tissues and their wide distribution suggests that they have a basic role to play in controlling

mineralization systems. Bone Gla-containing proteins are unique to bone and dentin and as such are expected to have a specific functional role to fulfill in these tissues.

Stages in the formation of bone are: (1) synthesis and extracellular assembly of the organic matrix framework, (2) mineralization of the framework, and (3) secondary mineralization as the bone constantly forms and reforms.

All of the salts and biomolecules associated with bone described above will play their own role(s) in the development of bone, the structure of which can vary considerably according to the location and use to which the resulting natural composite is to be used. Figure 2 shows pictorially

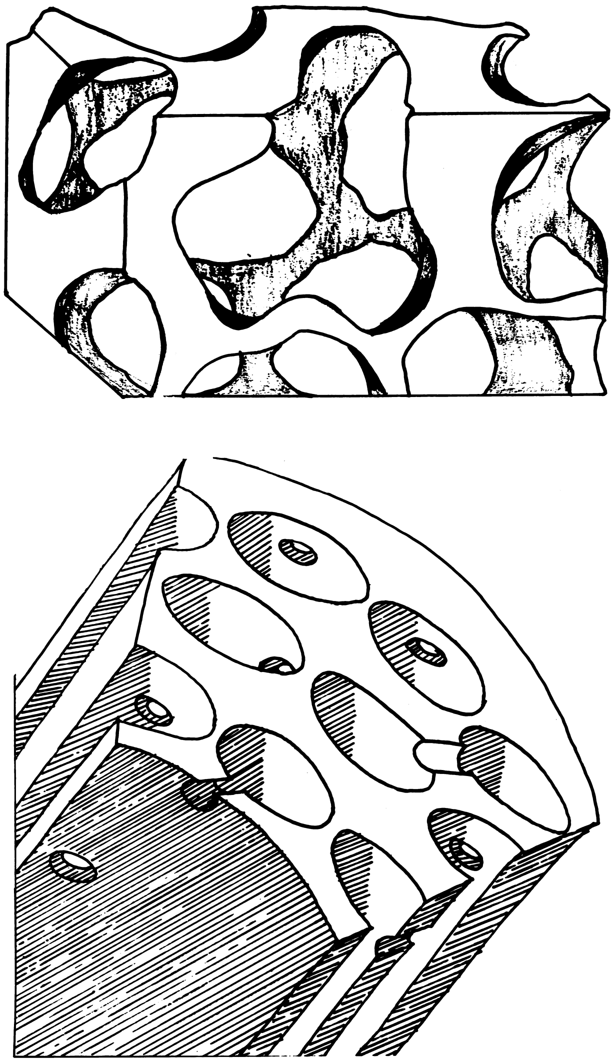


FIGURE 2 Schematic drawings of (a) human corticalbone and (b) humand cancellous bone. Note the difference in packing density and porosity between the two idealized structures. (Reprinted with permission from Perry, C. C. (1998). *In* "Chemistry of Advanced Materials." (L. V. Interrante and M. J. Hampden-Smith, eds.), pp. 499–562, Wiley VCH, New York.

the difference in porosity of cortical and cancellous bone with the former being found where load bearing is important. Bone is constantly in a state of dynamic equilibrium with its environment and changes with age. Changes with time will also be expected for diseased states and when foreign bodies (e.g., implants) are in close proximity to these phases although much less is known for such situations. An understanding of the structure and dynamics of natural materials should enable to design of materials for their replacement which will be chemically more compatible with those they are seeking to replace.

Mineralized cartilage contains much thinner fibers of collagen than are found in bone, high levels of water of hydration, and hydroxyapatite crystals, although there is no regular organization of the crystallites with respect to the collagen matrix.

Enamel is formed via the assembly of a matrix comprising both collagenous and noncollagenous proteins into which large oriented hydroxyapatite crystals are formed. The crystals may be of the order of 100 microns in length, 0.05 microns in diameter, and with an hexagonal cross-section. At maturity water and protein (including collagen) are removed from the tooth leaving a collagen-free composite.

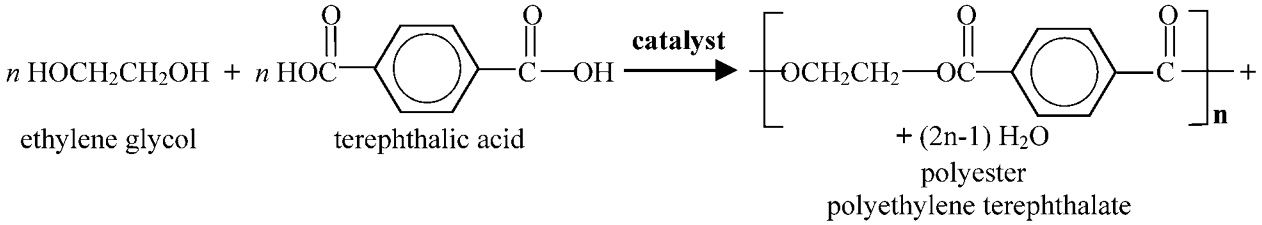
III. GENERAL REPAIR MECHANISMS AND BIOCOMPATIBILITY

Under normal circumstances most tissues in the body are able to repair themselves although the process and the presence or absence of scarring is tissue dependent. Bone repair occurs either through formation of membranous bone or through mineralization of cartilage. In facial bones, clavicle, mandible, and subperiosteal bones, membranous bone growth involves calcification of osteoid tissue (endochondral bone formation). In long bones the stages of repair include induction, inflammation, soft callus formation, callus calcification and remodeling.

Cartilage and skin can also repair themselves although scarring does occur. For skin, repair involves inflammation, immunity, blood clotting, platelet aggregation, fibrinolysis, and activation of complement and kinin systems. In the absence of a chronic inflammatory response, dermal wounds are repaired through deposition and remodeling of collagen to form scar tissue.

Enamel is not repaired by the body.

Early studies on biomaterials were based upon the idea that implants would not degrade in the human body or be involved in biological reactions. Hence the search was for bioinert materials, whatever their chemical composition. However, no material implanted in living tissue is inert and all materials elicit a response from the host tissue. For



SCHEME 1 Condensation polymerization.

toxic materials, the surrounding tissue dies. For nontoxic materials and those that dissolve, the surrounding tissue replaces the implant. For biologically nontoxic and inactive materials, a fibrous tissue capsule of variable thickness can form. For nontoxic and biologically active materials, an interfacial bond forms. Materials used in medical procedures should be *biocompatible* with the host tissue. Biocompatibility of an implant material is deemed optimal if it promotes the formation of normal tissue at its surface, and in addition, if a contiguous interface capable of supporting the loads that normally occur at the site of implantation is established.

Hence the current goal is to produce materials that are recognized and assimilated by the body.

IV. MATERIALS OF CONSTRUCTION

All medical implants are composed of polymers, metals, ceramics, or mixtures and composites of these materials. Tissue replacement with synthetic materials requires selection of a material or materials with physical properties most similar to those of the natural tissue (Table I).

A. Synthetic Polymers

These are the most widely used materials in health care and are used in almost all phases of medical and/or dental treatment. Typical polymers in use are listed in Table II together with selected uses.

1. Synthesis

Polymers are large molecules made from many smaller units called monomers that are chemically bonded to form the polymer. If only one species of monomer is used to build a macromolecule the product is termed an homopolymer, normally referred to as a polymer. If two types of monomer unit are used the material is known as a copolymer and if three different monomers are used then a terpolymer results.

Polymers may be formed by condensation reactions between complimentary functional groups to make poly(esters), poly(amides) and poly(urethanes) (Scheme 1).

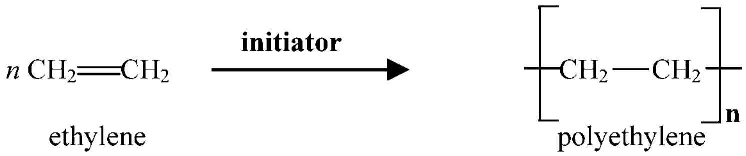
They may also be formed by free radical polymerisation of unsaturated compounds to give addition polymers (Scheme 2).

Examples of this class include poly(ethylene), poly(vinyl chloride) and poly (methyl methacrylate).

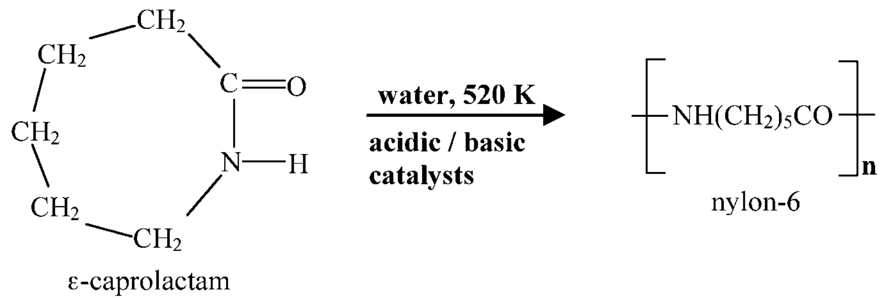
A third route to polymers includes ring opening reactions as in the formation of nylon-6 from ε-caprolactam (Scheme 3).

The structure, chemical functionality and physical properties of the polymer phase and hence its potential in medical devices depends on the monomers used and the method of synthesis. The presence of more than one functional monomer can lead to different arrangements of the monomers in the polymer chain (Fig. 3). These structural variations, including effects due to chirality have an effect on the way in which the polymer chains can pack together and hence they have an effect on the physical properties of the material itself.

Polymers are often compounded with a range of other materials to lower the cost and improve the performance of the proposed polymer device. Accelerators are used to increase the extent and rate at which cross-linking between the chains occurs. Antioxidants such as amines and hydroquinones are added to minimize the cracking of a device when exposed to oxidants. Fillers such as carbon black,



SCHEME 2 Addition polymerization.



SCHEME 3 Ring opening polymerization.

glass fiber, mica, and silica can be used to reinforce polymers and improve their mechanical properties. There are problems associated with the use of many of these additives because of their low molecular weight (in relation to the polymer phase) as they may be leached from the device causing deleterious effects on the device itself and on the human body. Not only can there be problems due to leaching of additives but also there may be problems due to the very components used in the synthesis of a material. For example, contaminants found in silicone polymers due to the synthesis route include; siloxane monomers, platinum catalysts, and peroxide derivatives. This problem is also significant in the production of bioceramics and composites for use in biomedical applications.

The mechanical properties of polymers are a consequence of the distribution and packing of the polymer chains (degree of crystallinity) and the transition temperature at which they change from a viscoelastic material to a rigid glass. They cover a wide range of strengths (as measured by values for the Young's modulus) and can therefore be used for a wide range of applications.

2. Polymer Modification

Polymers are required for use under aqueous conditions where the binding of proteins (for recognition by the body) and cells is required in order to produce a favourable biological response. Polystyrene, polyethylene and terphthalate, polyfluoroethylene and perfluorinated ethylene propylene copolymer are poor supports and require post-synthesis modification of their surfaces to improve biocompatibility. Gas plasma (glow discharge) methods have become popular as a way of fabricating a range of surface chemistries. Surfaces are produced with functionalities rich in oxygen groups including O, OH, surface sulfonate and carboxyl which are, in principle, more compatible with biological tissues than the carbon–carbon and carbon–hydrogen bonds present in the parent polymers.

It is also possible to use grafting (long used in industries based on polymers) to modify the surface chemistry of polymers. A grafted surface can be produced primaril by graft polymerization (often a free radical process) of monomers or the covalent coupling of existing

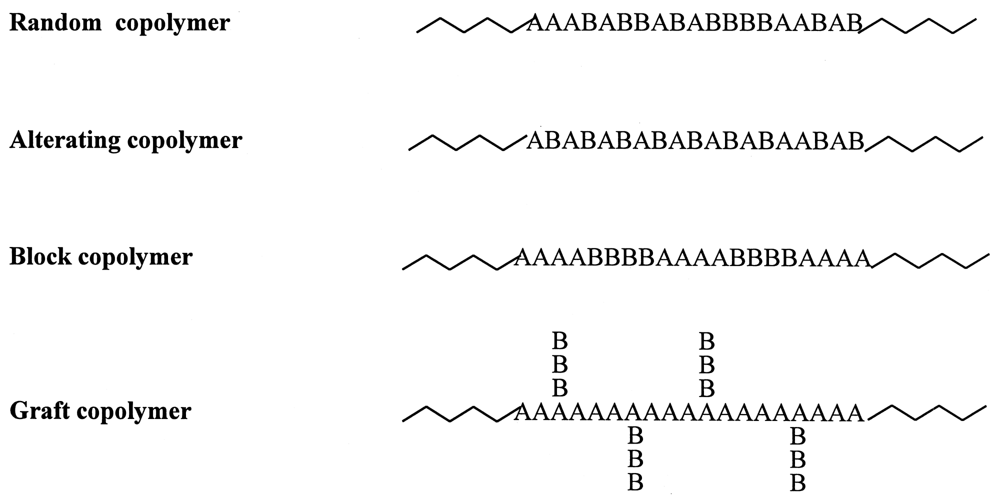


FIGURE 3 Schematic of polymer structures possible when two different monomers are used in the synthesis.

polymer molecules onto the substrate polymer surface. A terminal group reactive to functional groups present on the substrate polymer surface is required for the polymer chains to be used in the coupling reaction, while the graft polymerization method needs active species on the substrate polymer to initiate radical polymerization. An apparently grafted surface can also be produced by physical adsorption of existing polymers. The adsorbed chains are readily desorbed upon immersion in solvents or washing with surfactants unless the adsorptive force is remarkably strong. Polyethyleneoxide (PEO) has been used to modify artificial surfaces by the above chemical surface modification methods with the establishment of PEO-surface and PEO-interpenetrating networks. An alternative simple and cost-effective approach is to use the melt blend technique where small amounts of protein compatible additives, such as PEO, polyvinyl alcohol (PVA), poly(ethyl oxazoline) (PEOX), and poly(vinyl pyrrolidine) (PNVP) have been used to modify the substrate polymer. The base polymer could be a single polymer or mixture of ethylene–vinyl acetate (EVA), polypropylene (PP), glycol modified poly(ethylene terephthalate) (PETG), poly(methylmethacrylate) (PMMA), styrene–butadiene copolymer and polyamide-based copolymers. Materials modified by these techniques can show enhanced protein adhesion important for general biocompatibility or show reduced protein adsorption which is useful in the production of blood-contacting devices, chromatographic supports, coatings to minimize biofouling, separation membranes, contact lenses, immunoassays, protein drug-contacting materials, etc.

3. Biodegradable Polymers

Although there are many medical applications for polymers which require materials to be stable over a long time period, there are also devices such as sutures, small bone fixation devices, skin grafts, and drug delivery systems where polymers that break down in a controlled fashion are required. Materials that can be tailored to break down either unaided or by enzyme-assisted pathways under the conditions inherent in the location where the implant is found are desired.

Two main factors affect the biodegradability of polymers. The first is the chemical composition of the material and the presence of hydrolysable and/or oxidizable chemical groups in the main chain, suitable substituents in side chains, stereoconfiguration, balance of hydrophilicity and hydrophobicity, and conformational flexibility all contribute to the biodegradability of synthetic polymers. The second aspect is the morphology of the polymer sample with amorphous polymer regions degrading prior to crystalline and cross-linked regions. Functional groups such as amides, enamines, enol-ketones, esters, ureas, and urethanes when present in polymers all show some biodegradability.

Materials being developed include lactide/glycolide polymers, polyorthoesters, derivatives of pseudo- and poly-amino acids such as poly(l-glutamate), polyphosphazenes, poly($\in$-caprolactone) and tyrosine–polycarbonates. Advantages of these materials lies in their ease of processability and their biocompatibility with extracellular matrix components present in the body. They may be made as high surface area, porous devices thereby allowing cell movement and adhesion within the device both important for the assimilation of the material within the body and in the eventual replacement of the biodegradable component with the bodies own tissue.

Future applications of degradable polymers may involve their use in a foam form as scaffolds for the regeneration of specific tissues such as liver, bone, cartilage, or vascular walls.

B. Polymers from Natural Sources

In the search for biocompatible materials scientists have taken naturally occurring polymers and modifed the structure of the materials for a range of end uses. Materials based on collagen and hyaluronic acid are in common use for both dental, ophthalmological and maxillofacial reconstructive work where the “natural” polymer phase is used to fill out defects in the bone thereby providing a matrix around which bone will develop. Collagen is used in film, membrane, solution, gel, and sponge forms. It is also used in conjunction with glucosaminoglycans, tricalcium phosphate, hydroxyapatite, allogeneic bone, cells, and with drugs such as tetracycline.

Most implant procedures where collagen has been used have required the implant to be a permanent feature in the human body. The “collagen” used was usually partially degraded, enzyme-extracted collagen, or had been stabilized by cross-linking it with cytotoxic glutaraldehyde or chromium salts, or else had been assembled into non-natural polymeric structures such as films and sponges. An alternative approach is to maintain as much of the biological and organizational structure of collagen as possible using continuous collagen threads to make collagen fibers which can then be used to knit and weave fabrics with a structure more akin to that found naturally. Many different materials have been fabricated each with different bulk and extensibility properties.

Materials derived from hyaluronic acid are conventionally esterified (either 100% or fractional amounts) and the acid esters so produced have physicochemical properties which are significantly different from those of the parent molecule. Ethylhyaluronic acid ester (HYAFF-07)

TABLE IV Metals Used in Medical Devices

Metal	Medical device applications
Cobalt–chromium alloys	Dental applicances, fracture plates, heart valves, joint components, nails, screws
Titanium alloys	Conductive leads, joint components, pacemaker cases, nails, screws
Stainless steel	Fracture plates

and benzyl hyaluronic acid ester (HYAFF-11) are used to replace short nerve segments and in wound healing. The materials can be used as threads, films, fabrics and sponges and additional aplications are expected in plastic surgery and orthopedics.

C. Metals

Metals have a large range of applications as devices for fracture fixation, joint replacement, surgical instruments, external splints, braces and traction apparatus, as well as dental amalgams (Table IV). The high modulus and yield point coupled with the ductility of metals makes them suitable for bearing heavy loads without large deformations and permanent size changes. Metals are generally inert and if the composition is chosen carefully do not degrade in a saline environment. Metals are crystalline materials with a specific arrangement of metal atoms within a crystal lattice. Figure 4 shows the arrangements of atoms in the common crystal packing arrangements adopted by metals. The limits of a perfect crystal lattice are defined by grain boundaries where individual perfect crystals come together (Fig. 5). It is possible to incorporate a wide range of different atoms either within the crystal lattice, at interstitial sites within the crystal lattice (spaces/sites where atoms or ions are not normally located), or at grain boundaries and thus a multitude of metal-containing materials can be made. The deformation characteristics of a metal are determined by the grain size of the crystals as imperfections are concentrated at the grain boundaries. Mixing of metal atoms of different sizes as in the production of an alloy can serve to modify the properties of the metallic phase. Metallic elements used in the formation of implants include: aluminium (Al), cobalt (Co), chromium (Cr), copper (Cu), gold (Au), iridium (Ir), iron (Fe), manganese (Mn), molybdenum (Mo), nickel (Ni), niobium (Nb), palladium (Pd), platinum (Pt), silver (Ag), tantalum (Ta), tin (Sn), titanium (Ti), vanadium (V), tungsten (W), zinc (Zn) and zirconium (Zr). Nonmetallic elements that are used to modify the properties of the metallic phases include carbon (C), nitrogen (N), phosphorous (P), sulfur (S), and silicon (Si).

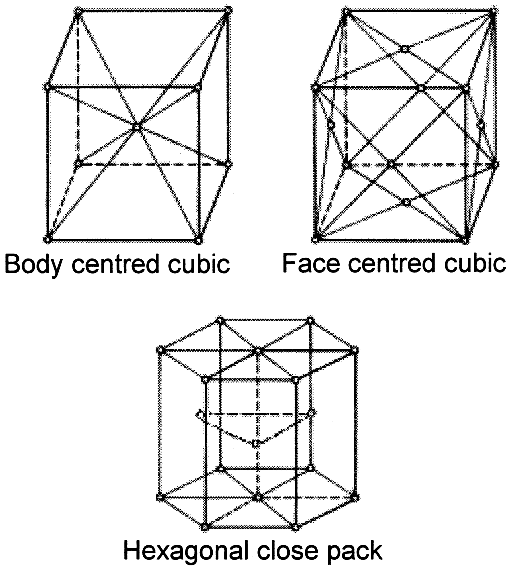


FIGURE 4 Common crystal packing arrangements for metals.

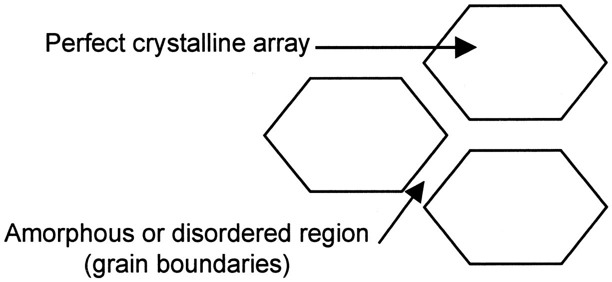


FIGURE 5 Schematic diagram of grain boundaries between crystallites. The boundaries may occupy only one row of atoms or more.

The metals and alloys that were originally used in the medical environment were selected on the basis of their strength and ductility although their original genesis may have been for a totally different purpose. To surmount problems due to corrosion under saline conditions alloys (homogeneous mixtures of the metallic elements at the atomic level) have been developed to render materials passive to corrosion. Another method is to promote the formation of an adherent oxide layer, often by acid treatment of the metallic phase. Both methods are currently used.

Although the properties of metals can be modified by the chemical addition of different atoms to the alloy mixture, physical treatments such as annealing, precipitation hardening, tempering, work hardening, and polishing can also modify the modulus, toughness, and surface properties of the metallic phase. Processing of the metallic material is necessary to produce functional components and a combination of brazing (complex part formed by heating in the presence of another metallic material), drawing

(wires and sheets formed from metallic ingots), forging (metallic forms obtained from dies), machining (for complex geometries), and welding (local heating and fusion to produce complex parts) may additionally modify the physical properties of the metal or alloy being used.

The materials currently used in the production of medical devices include stainless steels, cobalt-base alloys, titanium-base alloys, platinum-base alloys, and nickel–titanium alloys. Steels were the first modern metallic alloys to be used in orthopedics and initial problems with corrosion were overcome by modifying the composition of the steel with the addition of carbon, chromium, and molybdenum. Carbon was added at low concentrations (ca. 0.03–0.08%) to initiate carbide formation, while the addition of chromium (17–19%) facilitated the formation of a stable surface oxide layer and the presence of molybdenum (2.0–3.0%) was found to control corrosion. The compositions of stainless steels used can vary widely. Table V shows the limits for the chemical compositions of three different alloys containing eleven different elements together with the mechanical properties for the samples after annealing and cold working.

There are at least four compositions of cobalt-base alloys in use which are similarly designated by code numbers such as F75, F90, F562, and F563. Again, these differ in the relative composition of the following elements: manganese, silicon, chromium, nickel, molybdenum, carbon, iron, phosphorus, sulfur, tungsten, titanium, and cobalt. These alloys are used because of their superior strengths but they are difficult to machine and much more expensive to produce. Titanium-base alloys are found in many commercial medical devices and they are also used as coatings. For dental implants bone is found to grow best in the presence of titanium or materials coated with titanium where surface roughening during manufacture is also found to improve the performance. Platinum-base alloys are used primarily in electrodes for electrical stimulation for although they show excellent strength and corrosion resistance they are very expensive materials to produce and machine.

Figure 6 shows the clinical uses of metals in the human body. In many instances metallic implants have to be fixed in to the body and the implants must be compatible with the fixative which may be metallic (screws), ceramic (screws and other components) and/or polymer phases (e.g., glue). In the design of replacement components with high strength it is important that the compatibility of all of the biomedical components is required and investigated before novel implants are placed in the human body.

D. Ceramics

During the last 40 years a revolution in the use of ceramics has occurred. The revolution is the development of specially designed and fabricated ceramics, termed "bioceramics" when used in the body for the repair and reconstruction of diseased, damaged, and "worn out" parts

TABLE V Chemical Composition and Tensile Strength of Standard Stainless-Steel Alloys Used in Biomedical Applications

Composition	F55(%)		F138(%)		
Element	**Grade 1**	**Grade 2**	**Grade 1**	**Grade 2**	**F745(%)**
Carbon	0.08<	0.03<	0.08<	0.03<	0.06<
Manganese	2.0<	2.0<	2.0	2.0	2.0<
Phosphorus	0.03<	0.03<	0.025<	0.025<	0.045<
Sulfur	0.03<	0.03<	0.01<	0.01<	0.03<
Silicon	0.75<	0.75<	0.75<	0.75<	1.0
Chromium	17–19	17–19	17–19	17–19	17–19
Nickel	12–14	12–14	13–15.5	13–15.5	11–14
Molybdenum	2.0–3.0	2.0–3.0	2.0–3.0	2.0–3.0	2.0–3.0
Nitrogen	0.1<	0.1<	0.1<	0.1<	
Copper	0.5<	0.5<	0.5<	0.5<	
Iron	Balance	Balance	Balance	Balance	Balance
Ultimate tensile strength	**MPa**		**MPa**		**MPa**
Annealed	480–515		480–515		480>
Cold-worked	655–860		655–860		

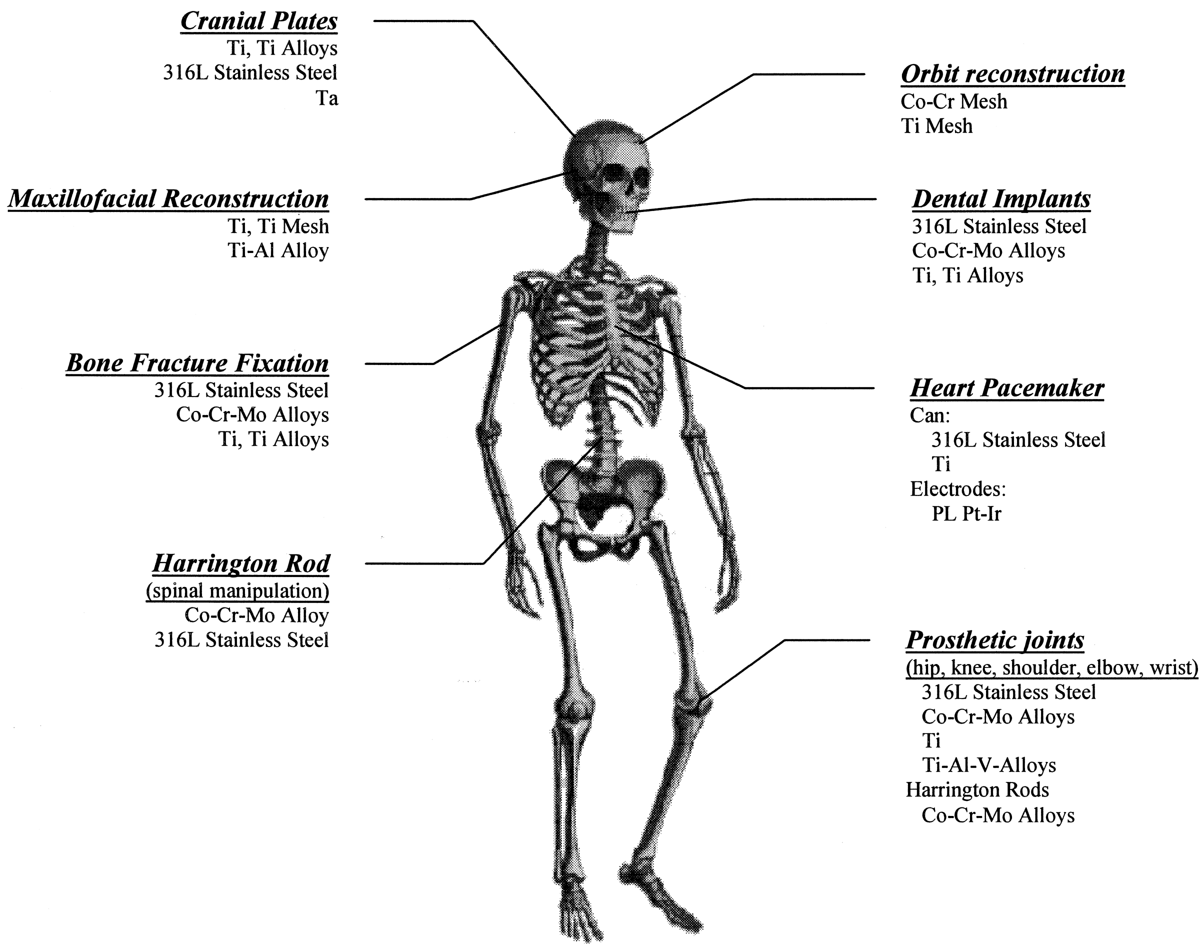


FIGURE 6 Clinical use of metals in the human body.

of the human body. Figure 7 shows many of the clinical applications of ceramics together with the materials used. Most applications relate to the repair of soft and hard tissues but ceramics are also used in the formation of replacement heart valves.

Ceramics include inorganic materials such as carbons, silica, and metal oxides in crystalline and glassy phases and salts such as calcium hydroxyapatite, found naturally in bone and enamel. Ceramics are stiff, brittle materials that are generally prepared by high temperature methods and the resulting materials are insoluble in water. The materials have high strength and hardness but they are only able to deform minimally under loading and therefore fracture easily. The materials are made in a variety of physical forms such as powders, coatings, and bulk phases. They can show a range of crystallinity from single crystals (e.g., sapphire), polycrystalline materials (e.g., alumina and hydroxyapatite), glasses (e.g., Bioglass®), glasses mixed with other ceramics (e.g., A/W glass–ceramic). They can also be used in polymer ceramic composites (e.g., polyethylene–hydroxyapatite). The specific material form used depends on the application. For example, single crystal sapphire is used as a dental implant because of its high strength. A/W glass–ceramic is used to replace vertebrae because it has high strength and bonds to bone. Bioactive glasses have low strength but bond rapidly to bone and are therefore used in the repair of bony defects. It should be noted in all applications, implants must remain within the human body for many years and the long-term stability of the materials used in the biological aqueous/saline environment must be considered.

1. Conventional Ceramics

These include oxides and salts of metallic elements and silicon together with carbons. The use of oxide phases in implants eliminates the problem of metallic corrosion in a saline environment with high-purity alumina (Al_2O_3)

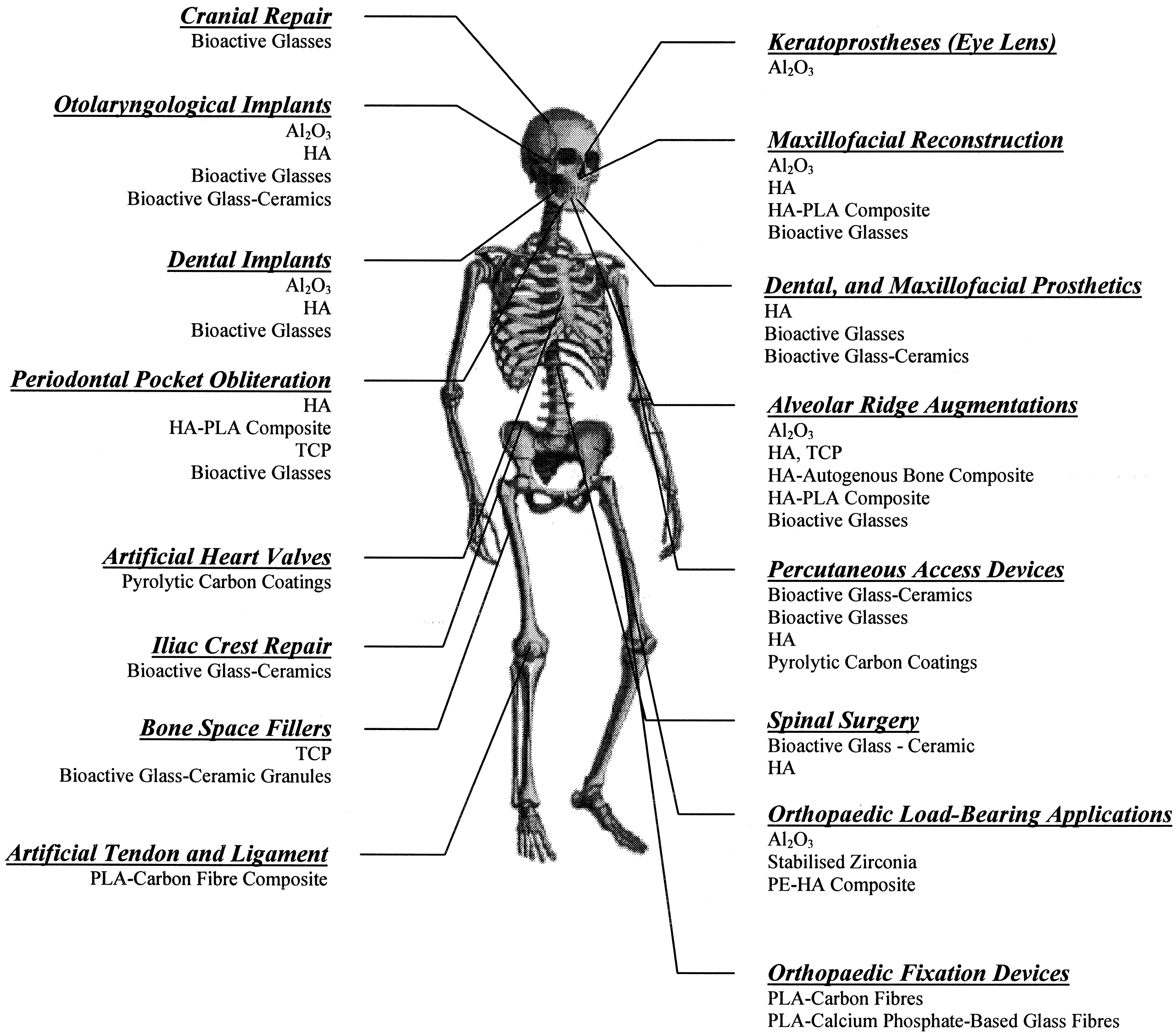


FIGURE 7 Clinical use of inorganic bioceramics in the human body.

being one of the first oxide materials to be developed for use in load-bearing orthopedic prostheses. Cermamics based on alumina mixed with beta-silicon nitride (β-Si_3N_4) known as Sialon® are also used with the mixture having higher fracture toughness than either component on its own. Zirconia (ZrO_2) and yttrium and magnesium stabilized zirconias are also used when applications require a fully densified material. Machinable glass–ceramics containing mica as the main crystal phase such as Macor® with a base glass composition of 47.2 SiO_2, 8.5 B_2O_3, 16.7 Al_2O_3, 14.5 MgO, 9.5 K_2O, and 6.3 F^-, and DICOR®, which contains significant amounts of the alkaline earth oxides are used in the production of dental crowns because the materials show excellent machinability, translucency, and good bending strength. BIOVERIT® II is a phlogopite-type glass–ceramic with two major crystal phases, mica and cordierite, present in the glassy ceramic. The translucency, mechanical properties, and thermal expansion coefficient can all be regulated in accordance with the required medical usuage. The material is biocompatible but not bioactive and is used in middle ear implants and in dental work. BIOVERIT® I has a different composition being a mica-apatite glass–ceramic and offers the possibility of a machinable glass together with bioactivity. Both of BIOVERIT® materials have successfully been used in head and neck surgery and in orthopedic surgery.

A range of carbons with different structure are used in the production of artificial heart valves and orthopedic applications. Glassy carbons (showing different degrees of short-range order) and pyrolytic carbons are both used. Glassy carbon manufactured by pyrolysis of a semi-coke retains some porosity and has relatively good wear resistance. Silicon carbide/carbon composites are prepared by impregnating a porous carbon with liquid silicon. Carbon-reinforced carbons and other reinforced carbons are made by impregnating organized filaments with a carbon filler, compressing, heat-treating, and carbonizing or graphitizing. Composites such as CFSiC admixed with SiC lead to materials with Young's moduli close to that of bone. They exhibit biological and mechanical stability and they are being investigated as new candidates for hip joint replacement rather than the metal or oxide phase conventionally used.

The other ceramic widely used are phosphate salts of calcium, with the chosen phase usually being hydroxyapatite. This material is conventionally prepared by thermal methods at temperatures well in excess of 1000°C. As a result of their preparation at high temperatures, the salts are carbonate free and are made up of much larger and more perfect crystals than those found in biological apatite minerals including bone. The imperfect crystalline structure of bone mineral leads to the natural material being soluble and reactive with respect to body fluids. In contrast, the synthetic materials are much less reactive than those found in living tissue and problems with biocompatibility can arise.

In all cases, solid nonporous implants do not allow for biofilm or cell attachment at any site other than the bulk surface. If porous implants of the previously mentioned materials can be made interfacial stability between the implant and tissue will increase as cells will migrate into the structure. For example, bone will grow in pores greater than 100 μm in diameter and a blood supply can be maintained throughout a material with such porosity. However, such materials show reduced strength and toughness. A compromise is the application of porous ceramic coatings to metals as in-growth of, for example, bone can occur at the porous interface with the mechanical load being carried by the bulk–metal substrate. Problems with such implants usually arise from any incompatibilities between the metal substrate and ceramic film rather than between the ceramic and the natural tissue which overgrows the implant.

2. Bioactive Ceramics

Bioactive ceramics are defined as those which are nontoxic and biologically active and that favor the development of an interfacial bond, 0.01 to 200 μm thick between implant and tissue. In all cases interfacial dissolution of the ceramic phase in the saline environment of the body leads to modifications in surface chemistry which affect the precipitation of the calcium phosphate phase prior to cell growth and the more general incorporation of the implant. Four major categories of materials have been developed. These include: Dense hydroxy(l)apatite (HA) ceramics, Bioactive glasses, Bioactive glass–ceramics, Bioactive composites.

Bioactive glasses are conventionally prepared by the traditional methods of mixing particles of the oxides or carbonates and then melting and homogenizing at temperatures of 1250–1400°C. The molten glass is cast into steel or graphite molds to make bulk implants. A final grind and polish are often required. Powdered materials are produced by grinding and sieving the ceramic to achieve the desired particle size characteristics. The chemical components of bioactive glasses include CaO, P_2O_5, Na_2O, and SiO_2. The bonding to bone has been associated with the formation of hydroxyapatite on the surface of the implant. Although a range of compositions can be used (up to 60% silica), an even narrower range of compositions are found to bond to soft tissues. A characteristic of the soft-tissue bonding compositions is the very rapid rate of hydroxyapatite formation. This has previously been attributed to the presence of Na_2O or other alkali cations in the glass composition which increases the solution pH at the implant-tissue interface and thereby enhances the precipitation and crystallisation of hydroxyapatite. The rate of hydroxyapatite formation has also been shown to be strongly dependent on the ratio of SiO_2, the glass network former to Na_2O, the network modifier in the glass. When the glass contains over 60% SiO_2 or more, bonding to tissues is no longer observed. The solubility and chemistry (including diffusion of Na^+ ions, for example, by the addition of La_2O_3) of the glass phase can be modified by the incorporation of other phases.

Problems which associated with the conventional high temperature method of production arise from:

1. Highly charged impurities such as Al^{3+}, Zr^{4+}, Sb^{3+}, Ti^{4+}, Ta^{5+} etc., which can be picked up at any stage of the preparation process. The incorporation of impurity ions leads to dramatic reductions in bioactivity.
2. Processing steps such as grinding, polishing etc. all expose the bioactive powder to potential contaminants.
3. There is a compositional limitation on materials prepared by the conventional high temperature methods due to the extremely high equilibrium liquidus temperature of SiO_2, 1713°C, and the

extremely high viscosity of silicate melts with high silica content.
4. High temperature processing leads to increased processing costs.
5. The relative expense of sol-gel materials relative to traditional precursors for melting gives increased processing costs.

An alternative route to the production of bioactive glasses is to use low temperature processing methods including "Sol-Gel Technology." Materials can be prepared from colloidal powders and simple molecular alkoxide precursors. The "glass" phase originally forms as a gel that can subsequently be dried by supercritical methods or conventional heating to moderate temperatures of 600–700°C. The two methods of solvent removal give a much better control over sample purity. Advantages of the technology are the ability to make materials with a much wider range of composition and/or microstructure through manipulation of the processing parameters. Figure 8 shows reaction pathway for the formation of a sol-gel material and application of the technology to the production of materials in a variety of different physical forms from films and coatings through to powders and glasses. All precursors show different reactivity in the initial stages leading to the formation of sol particles (1–500 nm diameter) and gel structure.

Using this technology it has been possible produce produce bioactive glasses from $CaO–P_2O_5–SiO_2$ mixtures. Materials produced by the sol-gel route show much

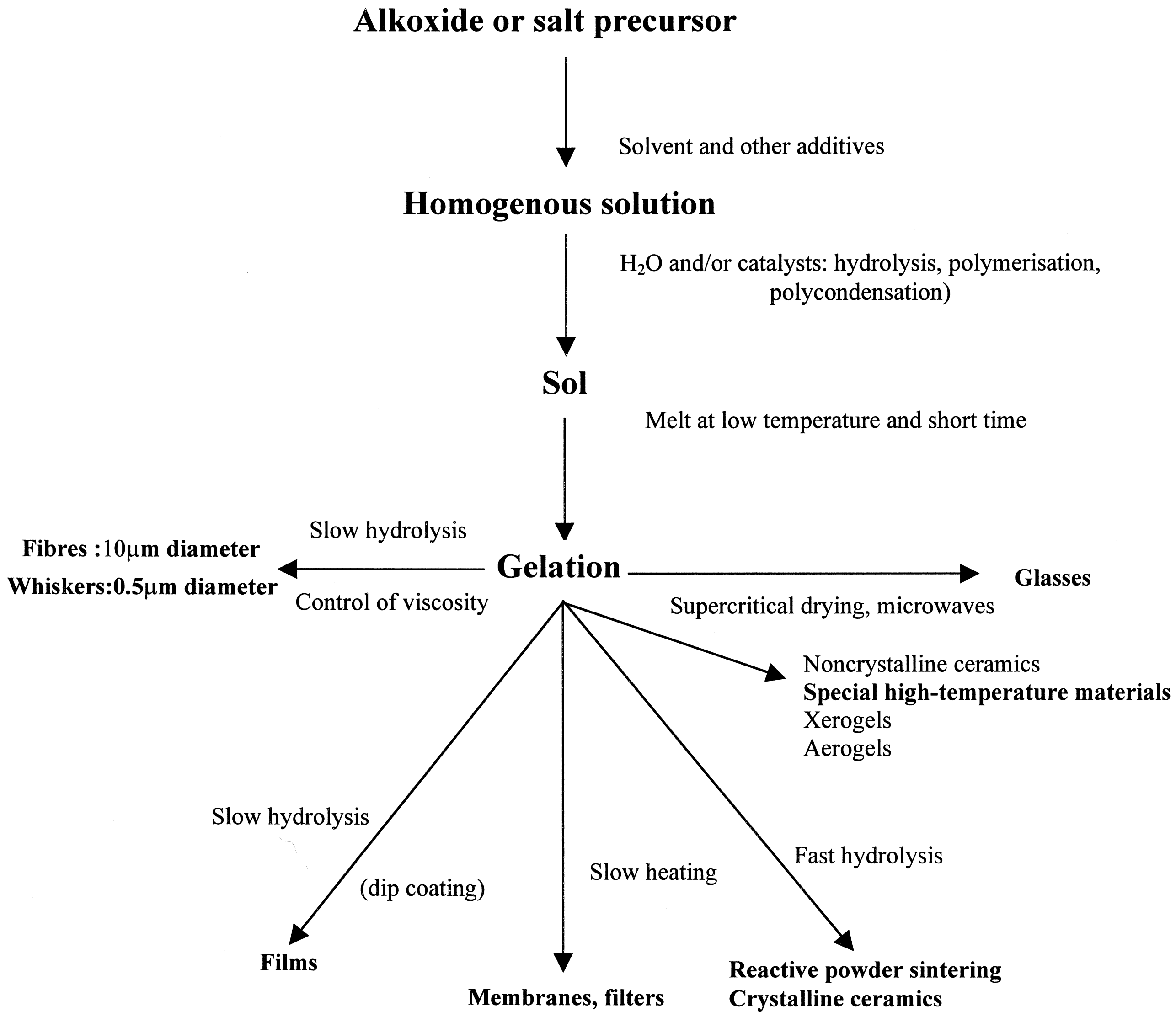


FIGURE 8 The sol-gel process for the preparation of materials.

higher levels of bioactivity for correspondingly lower levels of calcium oxide within the glass due to porosity generated during sample formation by this route which provides additional sites for mineral nucleation and sites for bone cells.

3. Toward Resorbable Implants

Biomaterials which are used to repair the body need to last as long as the patient does. At present this is not the case and some people may face several hip replacement operations, for example, each time there being less bone material (or less healthy bone material) for incorporation of devices. The current life expectancy of such replacements is on the order of 10 years at present. This needs to be doubled on tripled in the future. None of the materials described above is able to address the problem of tissue alteration with age and disease. The skeletal system has the capacity to repair itself, this ability diminishing with age and disease state of the material. The ideal solution to the problem is to use biomaterials to augment the body's own reparative process. Certain of the resorbable implants such as tricalcium phosphate and some bioactive glasses are based on this concept. Problems which exist with the development of resorbable materials are (a) the products of resorption must be compatible with cellular metabolic processes and (b) the rate of resorption must also be matched by the capacities of the body to process and transport the products of this process. In addition, as the material is resorbed and new material formed, the properties of both phases will alter and compatibility must be maintained at all times. This is difficult to achieve.

E. Composites

Composite materials are used clinically in order to take advantage of the desirable properties of each of the constituent materials while limiting the undesirable or deleterious properties of the individual phases. Composites cover a wide range of compositions and representative materials are listed in Table VI. Most of what is discussed in this section will relate to bioceramic composites. Bioceramic composites are either bioinert, bioactive or biodegradable. Examples of each of these classes of composite and their applications are given in Table VII.

The ceramic phase can be the reinforcing material, the matrix material or both. The incorporation of high strength fibers increases the mechanical strength of the composites while maintaining the bioactivity of the material. In the case of glass doped materials, the fracture toughness of the material increases dramatically and renders materials suitable for dental implantation and hip replacement therapy. The chemical composition of the fibers can be important in establishing continuity between the metallic component and the coating with titanium being an especially good candidate to achieve such an effect.

TABLE VI Composites Used in Medical Devices

Composite	Medical device applications
Glass, glass–ceramic or HAP with PMMA	Bone cement
Tricalcium phosphate/HAP with PE	Bone Substitutes
Glass–ceramic, quartz with BIS/GMA	Dental restorations
Drugs with various polymers/ceramics	Drug delivery
Carbon or glass fiber with PMMA and other matrices	All of the above applications but with increased strength and/or stiffness

1. Composites Based on HAP

There are many applications for calcium phosphate bioceramics can be seen in Fig. 7. The form of calcium phosphate used in orthopedic clinical applications is usually based on hydroxyapatite and β-tricalcium phosphate. The materials are widely used in composite formulations together with:

1. *Ceramics*. Mixed calcium phosphates, calcium sulfates, zinc calcium phosphates, aluminium calcium phosphates, metacalcium phosphates, sodium metacalcium phosphate, calcium carbonate, magnesium calcium carbonate and magnesium carbonate.
2. *Biological derivatives*. Bone derivatives (autografts, allografts and xenografts), collagen, dura, fibrin, amino acids, polyfunctional acids, inductive factors

TABLE VII Bioceramic Composites

Category	Examples
Inert	Carbon fiber reinforced carbon
	Carbon fiber polymeric matrix materials (polysulfone, poly(aryl) ether ketone
	Carbon fiber reinforced bone cement
Bioactive	A-W glass–ceramic
	Stainless steel fiber reinforced Bioglass®
	Titanium fiber reinforced bioactive glass
	Zirconia reinforced A-W glass–ceramic
	Calcium phosphate particle reinforced polyethylene
	Calcium phosphate fiber and particle reinforced bone cement
Resorbable	Calcium phosphate fiber reinforced polylactic acid

(bone morphogenic protein), growth factors (bone, epidermal tissue, cartilage, platelet, insulin).
3. *Therapeutic agents*. Hormones, antibiotics, chemotherapeutic drugs.
4. *Synthetic polymers*. Polylactic acid (PLA), polyglycolic acid (PGA), polycaprolactone (PCA), polyamino acids, polyethylene (PE) and high molecular weight derivatives, polysulfone, polyhydroxybutyrate.
5. *Metals*. Titanium-, cobalt-, and iron based alloys.

These materials can be developed in the form of particulates with a range of porosities, moldable forms, block forms, scaffolds, fibers, and coatings.

2. Bone Graft Materials

Bone graft substitutes are available based on alumina chemistry, silica, synthetic and natural calcium salts (phosphate, carbonate, sulfate, and hydroxide) and these materials combined with natural polymers such as collagen, and synthetic polymers such as PMMA, PHEMA, and UHMWPE. Both sintered and nonsintered materials based on calcium phosphate are available with the nonsintered versions showing greater biocompatibility (simply due to better resorption characteristics!). Materials can be produced with a high degree of porosity thus mimicking natural bone and allowing cells to permeate the implanted material over time. These can be produced from natural corals where the biomineralized skeleton of calcium carbonate is replaced by calcium phosphate. Examples include Interpre 200 and 500 with the materials being nearly nonresorbable. The same coral based materials, can be used in their calcium carbonate form without modification with resorption and replacement by fibro-osseus bone tissue. Calcium sulfate is routinely used as a casting material for fractures and is used for dental repairs together with porous hydroxyapatite granules (Hapset). The calcium sulphate is resorbed and can be replaced with the osseous tissue growing around the HA granules and holding them in place. Another biomaterial which makes use of calcium hydroxide together with PMMA coated with PHEMA is the hard tissue replacement polymer HTR. The implant material consists of PMMA beads which are sintered together to give a porous mass which is then coated with PHEMA and calcium hydroxide. The PHEMA coating absorbs a lot of water and a gel is formed at the surface containing calcium ions. This material is very biocompatible.

Other alternatives for implants are based on natural bone rather than the synthetic derivatives. For example, ashed bone can be used in combination with Ultra-High Molecular Weight Polyethylene, UHMWPE for coating on a porous implant for the purpose of biological fixation in joint repair. UHMWPE is the polymer of choice for the matrix material because of its abrasion-resistance, impact resistance, corrosion resistance and zero moisture absorption. The ashed bone provides the bioactive properties to prevent rejection.

V. THE WAY FORWARD, TISSUE ENGINEERING

Although many materials, both synthetic and natural, are used in medical devices and treatments it remains the goal of scientists and clinicians to be able to replace diseased or damaged tissues with living substitutes produced in the laboratory. These substitutes would be available in limitless quantities and be able to avoid rejection due to the body's own immune system. Tissue engineering is a way forward. Tissue engineering encompasses the study of cellular responses to materials implants, manipulation of the healing environment to control the structure of regenerated tissue, the production of cells and tissues for transplantation into the body, and the development of a quantitative understanding of biological energetics. Engineers, chemists, life scientists, and clinicians all have important roles to play in the furtherance of the discipline. Current areas of interest are the design of biocompatible casings for cell transplants, the development of polymer composites for patching wounds, the generation of scaffolds that guide and encourage cells to form tissues, the building of bioreactors for the large-scale production of therapeutic cells, and the establishment of experimental and mathematical models to predict cell behavior.

A. Prevention of Unwanted Tissue Interactions

Experimentation in this area began as early as 1933 with the use of synthetic nitrocellulose membranes to encompass cells and prevent an immune response. Current applications of the technology extend to the use of laboratory-grown skin in the treatment of burns and ulcers, the treatment of cancer patients via an increase in their marrow cells with culture external to the body and in the detoxification of liver cells from patients with liver failure.

Some materials that are being investigated as immunoprotective coatings are, alginate–polylysine coacervates, polyacrylates, polyphosphazenes, materials based on hyaluronic acid, and cellulose as well as hydrogel membranes directly synthesized on cells. The advantage of synthetic membranes as immunoprotective coatings is that they can be tailored for mechanical strength, biocompatibility, permeation characteristics, and

biostability via modifications to the synthesis procedure. They are being explored for the encapsulation of liver and dopamine-secreting cells for the treatment of liver disorders and nervous system disorders as well as the prevention of scar-tissue formation after surgery.

B. Directing Tissue Formation

Tissue engineering is causing significant advances to be made in "guided tissue formation." Isolated cells within a larger body of tissue exhibit little ability to organize themselves and form tissues. However, if cells are in fairly close proximity to one another they can grow and exhibit order in the formation of tissue with particular chemical and biological properties. The goal in tissue engineering has been to create an *in vitro* environment that would enable cells to organize themselves to form functioning tissues. The intention would then be to implant the artifically created cell structure to create new tissue or organs. In order to do this degradable polymer scaffolds have to be synthesized. Approaches to this have used copolymers of lactic acid and glycolic acid. Controlled porosity can be generated by use of salts and other additives that are later washed out prior to seeding the scaffold with cells. An alternative approach to the generation of porosity is to use carbon dioxide to dissolve the polymer phases and then allow them to reform around the gas bubbles on placement of the carbon dioxide dissolved polymer in an atmosphere of air.

Living structural supports hold cells close together using adhesive proteins such as fibronectin and vitronectin which bind reversibly to cell surfaces via a specific amino acid sequence-arginine–glycine–aspartic acid. This enables the cells to adhere to and interact with each other and with collagen and other constituents of the extracellular matrix. A modification of a lactic acid–lysine copolymer has been made with the tripeptide attached to the polymer via an amino group on the lysine. This approach to the synthesis of synthetic polymers with the essential components of natural proteins is being widely investigated. It combines the advantages of synthetic polymers with their desired materials characteristics such as strength, controlled degradation and processibility together with the essential features of biological polymers such as cell recognition and the capacity to control cell differentiation. A goal is to modify the specificity of the interaction between the polymer support and cells of choice. Specific advances in this area have been made for substrates used in bone repair which have been altered by binding peptides comprising lysine–arginine–serine–arginine to the surface. This allows specific interaction with osteoblasts rather than endothelial cells and fibroblasts that are also present in an implant area thus encouraging the formation of new bone.

Once a cell-seeded scaffold is implanted into host tissue it must allow blood vessels to grow into the new tissue. This can be promoted by the addition of a slow-release angiogenic growth factor which stimulates growth of new blood vessels. Although a polymer scaffold-cell-growth factor complex as described earlier was thought necessary for effective regeneration of tissue, scaffolds alone, cells alone, and diffusible bioactive factors alone may also serve to allow regeneration of tissue under specific circumstances. Tissues which are being prepared or are proposed for preparation via this route include cartilage, skin substitutes, and dental/orthopedic materials.

C. Large-Scale Culture of Therapeutic Cells

Conventional methods of cell culture have not been successful in the *in vitro* culture of cells for transplantation. Novel methods of culture are being devised in order to satisfy the demand for cultured cells. Special bioreactors and optimal, precisely controlled culture conditions are necessary to generate large quantities of therapeutic cells such as bone marrow cells for transplant to cancer patients undergoing chemotherapy. A goal for the future is the culture of stem cells from marrow which have been genetically modified to counter disease.

D. Future Directions

Tissue engineering has had successes in the blocking of unwanted rejection reactions between implants and host tissues, in the synthesis of polymer or polymer–cell composites for tissue repair without scarring, in the development of tissue cell culture for therapeutic cells and in the growth of simple tissues in the laboratory.

Future goals include the development of synthetic strategies to materials for implantation which overcome the bodies natural immune response and to generate "universal" donor cells that could be given to all as their immune characteristics would not be recognized by the host as "foreign." Learning how to stimulate the regeneration of complex multicellular structures *in vivo* is important and would allow the regeneration of (potentially) all tissues *in situ*. An important component of these advances will be the identification of specific cell signalling pathways and their spatial and time based involvement in the generation of organs built up from many tissue types. Practically all tissues are capable of being repaired by tissue engineering principles. Engineers, scientists, and clinicians need to use their understanding of synthetic and natural materials and the way in which the human body functions at the cellular level to develop the next generation of biomaterials and cellular transplants for use in the human body. The field is wide open for innovation in the development of a

new generation of biocompatible materials for use in the human body.

SEE ALSO THE FOLLOWING ARTICLES

Bioconjugate Chemistry • Bioinorganic Chemistry • Biomineralization and Biomimetic Materials • Biopolymers • Ceramics • Ceramics, Chemical Processing of • Polymer Processing • Tissue Engineering

BIBLIOGRAPHY

Burny, F., and Steinbuchel, A. (1998).“Biomaterials and Biodegradable Polymers and Macromolecules: Current Research and Industrial Applications,” Elsevier.

Combe, E. (1999). “Dental Biomaterials,” Kluwer Academic Publishers, New York.

Hubbell, J. A. (1995). “Biomaterials in tissue engineering,” *Biotechnology* **13,** 565–576.

Hubbell, J. A., and Langer, R. (1995). *Tissue Engineering, Chemical and Engineering News* 41–54.

Park, J. B., and Lake, R. S. (1992). “Biomaterials, an Introduction,” 2nd Ed. Plenum, New York.

Perry, C. C. (1998). “Biomaterials,” *In* “Chemistry of Advanced Materials” (L. V. Interrante and M. J. Hampden-Smith, eds.), pp. 499–562, Wiley VCH, New York.

Ratner, B. D. (eds.). (2000). “Biomaterials Science: An Introduction to Materials in Medicine,” 2nd Ed., Academic Press, New York.

Silver, F. H., (1994). “Biomaterials, Medical Devices and Tissue Engineering,” Chapman and Hall.

Silver, F. H., and Christiansen, D. L. (1999). “Biomaterials Science and Biocompatibility,” Springer-Verlag.

Williams, D. F. (1999). “The Williams Dictionary of Biomaterials,” Liverpool University Press.

Biomineralization and Biomimetic Materials

Paul Calvert
University of Arizona

I. Structural Biological Materials
II. Structural Proteins
III. Structural Polysaccharides
IV. Mineralized Tissues
V. Biological and Synthetic Processing
VI. The Process of Biomineralization
VII. Biomimetic Materials
VIII. Applications of Biomimetic Materials

GLOSSARY

Biomedical materials Materials for medical implants. May be soft, as in vascular grafts, or hard, as in hip replacement. May be wholly synthetic or of biological origin. Tissue-engineered implants are coated with cells previously removed from the patient and cultured on the implant.

Biomimetic (or bionspired) materials Synthetic materials formed using principles derived from the growth of biological tissue.

Biomineralization The process of forming inorganic material as part of the growth of biological tissue.

Freeform fabrication Family of methods to buiild objects from a three-dimensional computer representation driving some deposition system. As opposed to molding or machining.

Structural materials Materials with the principal function of carrying mechanical load, as opposed to sensors, actuators or electrical conductors.

THE MICROSTRUCTURE AND PROPERTIES of biological tissues are of interest because they influence the morphology and behavior of each species. Biological tissues are also of interest in that they can suggest new designs for synthetic materials and new ways of using synthetic materials to achieve an engineering goal. In surgery, we have the more difficult problem of developing synthetic materials that will work in tandem with animal tissues. The following article contrasts synthetic and biological approaches to processing, to material structure and to the use of materials in design, especially as this applies to mineral-reinforced composite structures.

A key observation concerning biological materials is that they are all composites. At any scale above 10 microns there are no uniform structures in biology. This makes the concept of a "biological material" rather uncertain because structures and properties change with position in the body, with the individual plant or animal, and with time. Thus, we can consider "bone" to be a material, but then distinguish the structures of woven, Haversian, and lamellar bone. We then subdivide the bone into dense and cancellous. Finally, the details of the structure will vary with site. Also, in contrast to assembled machines, sharp boundaries between structural materials are rare so it is quite difficult to define where bone stops and mineralized cartilage or mineralized tendon starts. For these reasons, it is also relatively difficult to characterize tissues and compare their properties with those of plastics, ceramics, or metals. This article will summarize biological polymer matrix materials, discuss mineralized tissues, and then discuss biomimetic composites and ceramics.

I. STRUCTURAL BIOLOGICAL MATERIALS

In their survey of biological materials, Wainwright *et al.* (1986) make a division into tensile materials, rigid materials, and pliant materials. Following a more conventional materials division, we will discuss polymers, ceramics and mineralized polymers, and gels; however, such boundaries are even less clear in biology than in the synthetic world.

Two processing-induced limitations should be recognized for structural polymers in biology. First, they are all formed from aqueous solution and so are all very sensitive to plasticization by moisture. In most cases, the properties of the dry material have little relevance because they will only occur in a dead organism. The plasticizing effect of water on biological polymers can be regarded as parallel to the softening effect of increased temperature on synthetic amorphous polymers. As temperature will take a hard polymer through the glass transition into a rubbery state, so will increased water content convert amorphous proteins from glass to rubbery. There is little sense in measuring the mechanical properties of biological materials without defining the water content.

Second, the growth process allows a variety of routes to the formation of fibers but there are no simple ways to form a dense isotropic plastic. Thus, even essentially isotropic materials will have a fibrous composite microstructure. Weiner *et al.* (2000) have argued that many biological structures can be viewed as a search for isotropic properties, or at least orthotropic properties (strong in two dimensions), from fibrous materials. This would reflect the unpredictability of stresses encountered by a structure in a dynamic environment.

II. STRUCTURAL PROTEINS

Silks have been reviewed Kaplan *et al.* (1997). Many insects and arachnids make extensive use of these protein fibers for a range of purposes. Spiders typically produce five different silks for the radial and spiral parts of the web, for a sticky web coating, for wrapping prey, and for the dragline. The properties variations come from differences in the amino acid sequence of the polymer. Silks are stored in a gland as an aqueous solution, at least some of the time in a lyotropic liquid crystalline state. For the strong silks, shear at the spinneret leads to a change in conformation to a very highly oriented and stiff fiber. The combination of strength, stiffness, and toughness shown by spider dragline silks has led to efforts to characterize, clone, and produce a synthetic version of this material .

Bulk crystalline polymers, such as nylon or polyethylene, have an amorphous fraction of 30 to 50% that arises from the inability of entangled polymer chains to become completely ordered. This crystallinity can be controlled to some extent by processing and can be reduced by copolymerization to introduce random irregularity into the chain. Fiber structures are less easy to resolve but behave similarly. In many silks, the structure seems to be blocky, with irregular sections spaced along the chain to define noncrystallizable sections. We are not yet able to resolve the role of these irregular sections in stabilizing the liquid crystalline state and in defining the final structure and properties of the fiber. In all biological polymers, there is a degree of control over the molecular structure that may be very important for properties and cannot be duplicated synthetically.

Collagen is the structural material of skeletal animals and has properties that are quite inferior to those of cellulose, chitin, or silk. The key to its use seems to lie in the versatile processability. Soluble procollagen is formed in the cell and exported into the growing tissue. An enzyme cleaves a bulky end section from the molecule to form tropocollagen that organizes into a triple helical structure. These triple helices self-assemble into collagen fibrils that make up the bulk of tensile structures such as tendon and ligament. As it ages, the collagen becomes cross-linked, which increases the stiffness and strength but reduces the toughness. Baer and co-workers (1991) have discussed the structure and properties of collagen in tendon and ligament.

Keratin can be seen as a biological answer to the need for a tough plastic equivalent to nylon. As hair and fur, it is a fiber. As epidermis, it is a film, and as hoof or horn, it is tough solid. The structure contains fibrils, which are built from three-stranded ropes of alpha-helical chains in a coiled-coil arrangement. The fibrils are embedded in a cross-linked matrix of amorphous protein that is heavily

cross-linked by cystines (sulfur–sulfur links). The high sulfur content is a defining characteristic of members of the keratin family and gives burning hair its characteristic smell. The fibrils confer great toughness by converting to a more extended beta-sheet structure under stress. Feather keratin is a beta-sheet structure presumably for providing higher stiffness. Amphibian skin is also a beta-sheet, again presumably to limit water swelling. The best-studied example of keratin is wool, but there is also recent work on hoof keratin.

III. STRUCTURAL POLYSACCHARIDES

Vincent (1980) discusses insect cuticle, which is a composite of chitin fibers embedded in a cross-linked protein matrix. Chitin is a polysaccharide, similar to cellulose but with acetylamino substitution. Metabolically, polysaccharides should be less expensive than protein since nitrogen has a much lower abundance in the biosphere than carbon, hydrogen, and oxygen. It is not clear what improvement in properties or processing led the insects to select chitin in the place of cellulose as their main structural material. The layered structures of cuticle, with various sequences of fiber orientation, do strongly resemble the layered structures of carbon-fiber composite laminates. Given that both insects and military aircraft are lightweight, roughly cylindrical systems, the resemblance cannot be accidental. Gunderson and Schiavone (1995) have discussed how insects adopt unbalanced or asymmetric layups that would not be used in synthetic composites. Some of the patterns also involve thick layers oriented along the major structural axes, with many fine layers forming the rotation from one direction to the next, apparently to delocalize shear stresses that would cause delamination.

Many layered biological systems resemble cholesteric liquid crystals in the rotation of orientation from layer to layer. This has raised recent interest in whether they actually are liquid crystals, in that the rotation forms spontaneously as a result of interactions between fibers in successive layers . This would occur in a fluid state, which is subsequently embedded in a hard matrix. The core question is really whether the rotation pattern is directly controlled by some form of oriented extrusion during the deposition process or is controlled through the surface chemistry of fibrils deposited in successive layers.

The mechanical properties of chitin are difficult to define because large oriented samples are unavailable. In cellulose, many plants contain bast fibers with very highly aligned polymer, allowing us to calculate the stiffness of the polymer and so analyze the properties of wood and other plant materials. The stiffness of cellulose in such fibers, 40 GPa wet and 100 GPa dry, is comparable to the stiffest synthetic fibers, such as Kevlar®. The polymer chains are wholly aligned with the fiber axis.

Plant cells are hollow tubes with spirally wound fibrils. Successive layers go in opposite senses, making a crisscross pattern. The outer and inner layers may be wound at different angles to the rest. This is very like the winding of fibers on a composite pressure vessel, such as a pressurized gas tank. The wood composite structure is designed to retain internal pressure or longitudinal compression with very high energy absorption on fracture as the windings collapse inwards.

Plants again have the problem of how to manufacture the polymer without it enveloping and choking the production site. It is laid down by a "track-laying" system, which seems to carry out the polymerization at the cell surface on a moving organelle and which draws glucose for polymerization through the membrane from inside.

In wood, hemicelluloses and lignin act as a matrix bonding the cellulose fibers into a composite structure. It is a puzzle that the stiffest, strongest, and cheapest of the biological polymers is not used at all in animals.

IV. MINERALIZED TISSUES

In the synthetic world, some applications involve predominantly tensile loading, the wall of a pressure vessel being one example. Much more commonly, parts will be loaded in compression or bending. While strong fibers in the form of a rope can provide excellent resistance to tension, they are of little use in compression. Large animals with an external or internal skeleton will need stiff materials that are more isotropic in their properties and so can withstand the varying stresses that come from moving around and colliding with other objects. Mineralized composites provide improved compressive properties over purely polymer structures. Compared to stiff polymers, minerals can also be formed at a lower metabolic cost for a given level of stiffness, but do increase the overall weight. A wide range of minerals is found in microbes, plants, and animals but silica, calcium carbonate, and hydroxyapatite are the most important.

A. Silica

Silica occurs as spicules (short reinforcing rods) in sponges. The spicules are typically 10 μm in diameter and 100 μm long and may be simple rods or complex branched structures. They apparently form by aggregation of silica nanoparticles onto a thread of polysaccharide or protein within a vesicle, a membrane-enclosed space inside the organism. The material is amorphous, highly hydrated, and not fully dense. There is one example of a deep-sea

sponge, *Monoraphis*, that is attached to the seabed by a large silica rod, 1 m long. The structure of this material is a series of concentric layers where the weaker interlayer regions may enhance the toughness of the structure, like a laminated glass windshield.

Siliceous diatoms are single-celled prokaryotes, about 100 μm in diameter, enclosed by a porous silica shell, in the form of two overlapping dishes, like a petri dish. The organism multiplies by dividing into two disc-shaped cells and then forming two new dishes back-to-back. The cells then separate, each with one new half shell and one old one. Recent work on the proteins involved in silica mineralization has started to clarify the details of silica deposition in diatoms and sponges. Plants often contain silica as a reinforcing material, and the structure of bamboo has been much studied in this context. Seawater and soil water contain low levels of dissolved silica as silicic acid, which the cell probably binds and transports as a complex with catechols (dihydroxybenzene) and then converts this to silica nanoparticles.

B. Carbonate

Calcium carbonate is widespread as a protective shell in marine animals, from single-celled coccoliths through coral, gastropods, and bivalves. It also occurs as a reinforcement in the cuticle of crustaceans such as crabs. The use of calcium carbonate for shells, rather than silica, may reflect the greater control of structure available through calcium-binding proteins and through control of crystal morphology with nucleators and growth inhibitors. Silica does occur as a component of many mollusk teeth, as do various iron oxides.

There has been much recent work on the formation, structure, and properties of mollusk shell. Most shells are constructed of calcite and aragonite in various arrangements with small amounts of protein, up to about 5% by weight. Higher protein levels would be expected to increase the toughness of the shell but at the expense of reduced stiffness. It can be assumed that particular shell structures are adapted to the particular lifestyles of the animals and the stresses encountered. For instance, it might be expected that a swimming bivalve, such as a scallop, would have a stiffer, lighter shell structure. One of the most studied shell structures is nacre, or mother-of-pearl, which is a layered structure of aragonite plates, 0.5 μm thick by several microns wide. Between each layer of plates is a thin protein sheet that is responsible for nucleation of the aragonite layers and acts as a crack stopper, which gives wet nacre a very high fracture energy.

While the details of shell formation are not understood, it is clear that organic content contains proteins capable of selectively nucleating calcite or aragonite in a specific orientation and proteins that inhibit growth of specific crystal faces. These work in concert with the supply of calcium and carbonate from the mantle tissue. The mantle adds new shell at the existing edge by extruding over the lip to extend the outer and inner faces.

The vaterite form of calcium carbonate often occurs in laboratory crystallization and is occasionally found in shells. There is increasing evidence for precipitation of amorphous, hydrated calcium carbonate as a metastable precursor for the crystal. In some cases, including lobster cuticle and a sponge, stable amorphous calcium carbonate is found.

C. Hydroxyapatite

The human body is supersaturated for both calcium carbonate and for hydroxyapatite, $Ca_{10}(OH)_2(PO_4)_6$, the mineral of bone and tooth. Since the environmental availability of phosphorus is limited, its use as a reinforcement seems peculiar. Possibly bone acts a phosphorus reservoir. Bone is essentially a polymer-matrix composite reinforced with ribbons of hydroxyapatite. The structure and properties of bone are discussed by Currey *et al.* (1995). A key current question is the way the lamellar structure gives rise to a combination of high stiffness, high strength, and high toughness. Similar synthetic composites tend to break at relatively low strain and so have low toughness.

Dentine has a structure similar to bone but with significant differences in the mechanism of mineralization. Tooth enamel is almost wholly mineral with a fibrous structure that forms under the control of a completely different set of mineralization proteins, including amelogenin.

V. BIOLOGICAL AND SYNTHETIC PROCESSING

In practical engineering, where cost is very important, there is little sense in designing a part without a very good idea of how it will actually be made. Materials selections are then made once the manufacturing route and basic design have been decided. The same must be true in biology; many interesting structural features may be primarily a consequence of the growth process and only secondarily a source of improved properties.

Small numbers of synthetic parts can be made by subtractive processes such as machining but the process is slow and wasteful. Molding is generally cheaper and faster. Hot liquid material is injected into a hollow tool and allowed to solidify. Extrusion, forging, rolling, and other hot processes can be viewed as variants on molding. The sintering of ceramics is a separate case; compacted powder is heated until surface energy drives the slow shrinkage and densification to a solid. Chemical processing is not common in the synthetic world. Thermosetting resins such as

epoxies are solidified by chemical reaction and there is some exotic reactive processing of ceramics. Generally, chemistry has the problem that small changes in starting conditions or purity can cause big changes in the reaction kinetics, so reliability is poor.

In biology, all processing is essentially chemical. Soluble reagents are fed to a site where they combine to make a solid plus dissolved by-products. In the case of epoxies, mentioned above, there is little volume change between the liquid and solid states. In forming solids from solution, there is also a massive shrinkage to be accommodated. The biological equivalent of molding can be seen in the formation of isolated particles within an envelope of lipid membrane. Examples include the formation of silica in sponge spicules and diatoms and calcite in coccolith skeletons. The wall of the "mold" is now permeable to allow reagents in and soluble products out. This approach is suitable for isolated particles, such as the magnetic particles in magnetotactic bacteria and for the wall of a single-celled diatom. The particles may also be later assembled into a simple framework, as in the sponges. However, this method does not lend itself to strong, dense, load-bearing structures suitable for large plants and animals.

A problem with building a large solid object by chemical precipitation is to avoid surrounding and entombing the cellular machinery for providing the reagents and removing the products. The obvious solution is to build the solid layer-by-layer, such that a layer of cells provides material to add to the surface of the growing solid and retreat ahead of it. The layer approach is apparent in the growth rings of trees. Bone, tooth, and shell form in the same way. In wood and skin, it is the cells themselves that become the structural unit by depositing solid cellulose or keratin in or on the cell. Each new layer starts as a new layer of cells.

In bone and tooth, the deposition is external to the cells. In the case of bone growth, the cell layer continuously forms new collagenous matrix material. This then mineralizes and converts to hard bone as new matrix is deposited over it. The control system thus promotes crystal growth in layers several microns from the cell surface while not mineralizing the freshly formed matrix. Many proteins are known to be associated with this process control but it is not yet clear how they work together.

One natural consequence of this layerwise growth process is that layers of different material can readily be formed within a single solid. As will be seen below, biology makes extensive use of layered structures to add toughness to strong materials. Recent developments in ceramics have also focused on the use of layered structures to add toughness. New methods of freeform fabrication should ultimately allow the production of complex layered structures resembling those of biological materials.

A result of chemical precipitation is that it is relatively difficult to make dense structures. The large change in volume as a soluble polymer assembles into fibers or as a mineral precipitates will lead to highly porous structures unless the deposition rate is extremely slow. Any structure forming in a diffusion field will tend to grow towards the source rather than filling in gaps in a layer. Very slow growth will allow equilibration, which will favor dense structures. One solution to this is to lay down a mesh of strong fibers, fill in the pores with a softer matrix, and then arrange for this to slowly expel water and harden chemically. This kind of process is seen in wood and in cuticle and very similar issues occur in the formation of carbon–carbon composites. For such reasons, biological materials must be composites.

VI. THE PROCESS OF BIOMINERALIZATION

It has long been recognized that most biological tissues mineralize by precipitation in an existing organic matrix that controls crystal form, size, and orientation. Originally, it was thought that the key step would be nucleation on a protein where a suitable spacing of charged groups would match the crystal lattice to be formed. This picture was supported by the observation that acidic proteins extracted from mollusk shell would inhibit crystal growth of calcium carbonate from a protein solution but would nucleate crystal growth if the protein were immobilized on a surface.

Since then the picture has become much more complicated. Studies on nucleation at Langmuir monolayers and on self-assembled monolayers have shown that surface ionic charge is important but there is no strong evidence for lattice matching. While the nucleation effect can be important in growing surface-attached mineral films, it seems to work only in a window of concentration just below that at which nucleation occurs readily in solution. This window can sometimes be widened by adding growth inhibitor to the solution. The films that form often seem to be limited in thickness, contrary to the reasonable expectation that growth would readily continue once a mineral layer had covered the substrate. In the case of molluscan nacre it has been shown that some species, at least, nucleate each new layer of aragonite by growing through holes in a protein layer, rather than via a nucleating protein as had been thought.

Studies of protein synthesis during biological mineralization have shown that large numbers of proteins are being produced during the growth of tooth or bone. This gives us the problem of too much information, as it is not at all clear why so many are needed and what they all do. It is clear that much of the control is via inhibition of growth on specific crystal faces, leading to control of crystal shape and orientation. At first it might seem strange that a protein can bind so specifically to one crystal face, but a difference in binding strength may be all that is needed to change the relative growth rates of two crystal faces

and this will lead to a change in crystal habit. One essence of control is that signals can be turned on and off. Proteolytic enzymes which degrade specific inhibitors can also be expected to be part of a mineralization process.

Studies have also demonstrated that amorphous calcium carbonate does occur in some species and is a transient mineral in some others. This structure presumably results from high levels of incorporated protein in the structure. Other work on solution growth has shown that metastable complexes of calcium and acidic polypeptide can be important in growth of carbonate films. It has long been known that silica species in solution will promote precipitation of hydroxyapatite on many substrates. Again, some metastable complex is presumably involved. De Guire *et al.* (1998) have studied biomimetic growth of various minerals on substrates treated with self-assembled monolayers. They have shown that some cases seem to correspond to growth by addition of individual ions, while others involve colloidal attraction of preformed clusters to the surface and aggregation. Matijevic has shown that apparently crystalline, faceted particles of many minerals formed from dilute solution are actually aggregates of clusters. Thus, many mineralization processes in complex solutions may involve an intermediate cluster or polymeric state.

One striking example of the complexity of biomineralization is the fact that bone is largely mineralized by hydroxyapatite entrained in collagen fibrils, where it is believed to nucleate at the acidic terminal regions of collagen triple helices. Many synthetic studies have produced the mineral on collagen fibrils but none has produced mineral in the fibrils. We are missing some key aspect of the process.

It should also be kept in mind that mineralization is a process that occurs in space and time. Mineralized tissue is generally formed by a layer of cells that sequentially deposit organic matrix and mineral and move back. There is a structural gradient away from the cell surface; some species will act locally while others must diffuse some way to their site of action. In bone formation, matrix vesicles also provide the cells with the possibility of delivering species into the mineralizing zone, several microns from the cell surface. In mollusk shell, day–night cycles may also provide the structural sequence. Studies of "flat pearl" have also shown the progression of structure formation on a glass surface embedded under the mantle of a mollusk. Most studies of biomimetic mineralization have used constant precipitation conditions.

VII. BIOMIMETIC MATERIALS

The concept of biomimesis has long been used in chemistry in the context of compounds with enzyme-like catalytic action. Since the mid-1980s it has been applied to materials, particularly with a view to producing ceramic and composite materials with improved toughness, analogous to shell, tooth, and bone. There was some discomfort in the materials community with the idea that we wanted to mimic biological materials, in the sense of producing an indistinguishable copy. However, mimesis more generally refers to copying some essential aspects of a thing rather than duplicating or faking it. The phrase "bioinspired materials" is also used to express the idea in more familiar terms but is grammatically less desirable as it is a Greek/Latin hybrid. The field has now spread to include a group of loosely linked goals in new materials and processes, which are surveyed below.

It should be kept in mind that materials development is considerably upstream from the development of new products. It is quite typical for new materials to find their way into commercial products about 20 years after their discovery. Examples include Kevlar®, high-Tc ceramic superconductors, piezoelectric polymers, and gallium arsenide. It is also typical for the first applications to be quite different from those originally proposed and for their impact to be modest compared to that suggested during the initial excitement. A personal view of the status and prospects for biomimetic materials will be given at the end of this article.

A. Polymers

The core difference between the proteins and synthetic polymers is that protein synthesis provides total control of the sequence of units along a chain while the best-controlled polymerizations can only provide several blocks of functional units on a chain. One obvious goal is the formation of synthetic polymers with enzyme-like catalytic activity. Many enzyme-active sites have an array of active groups held in close proximity so as to interact with the substrate (target molecule) and to reduce the activation energy for reaction by a precise spatial array of ionic or hydrogen-bonding interactions. To achieve such a precise spacing of active groups on a synthetic polymer would require a rigid structure, which would in turn normally render the material insoluble and so inactive. Recent studies of dendrimer molecules with highly branched structures may lead us to the required combination of flexible and soluble outer structures combined with a highly structured core. Many proteins also go through large changes in shape in response to binding of substrates or other energy inputs. This seems to require a structure where two well-defined conformations are closely balanced in energy so the molecule can flip from one to another. Such a change is not likely in a wholly flexible system but again some subtle combination of flexible segments and rigid units is required. We still have much to learn in these aspects of macromolecular design but are acquiring both the synthetic tools and the understanding.

As revealed by studies of silk, structural proteins include regular and random regions in the chain structure which would be expected to give rise to crystalline and amorphous material. In contrast, the amorphous component of synthetic polymers arises from entanglements of the coiled chains that cannot be resolved during the crystallization process. New polymer properties may be achievable once we can design polymer chains with such controlled sequences.

There has been a report of enzyme-like activity in a block copolypeptide, which enhances the rate of hydrolysis of tetraethoxysilane (TEOS, a standard reagent in sol-gel chemistry) as a suspension in water. If this block structure is a sequence of units of a hydrophilic amino followed by units of a hydrophobic amino acid, it would be expected to be active at a water–solvent interface. The morphology of the silica that forms is dependent on the structure of the copolymer. This system is biomimetic both in the sense of employing a polypeptide catalyst and in the sense of it functioning in a multiphase system, since biological processes rarely occur in homogeneous solutions.

Many tissues, such as cartilage, arterial wall, and the walls of soft marine organisms, are swollen polymer structures. Swollen polymers, such as plasticized polyvinylchloride, do occur in artificial structures but they are usually avoided because loss of plasticizer leads to shrinkage and cracking. Even the swelling of wood with changes in humidity is a major impediment to its use in structures, though here years of experience have taught us how to design around it. Skin does change in volume and properties as it takes up or loses water. The structure of amphibian skin keratin is apparently different from mammalian keratin for this reason.

There have been many suggestions that designers should make more use of soft structures. These could take the form of composites of hard fibers with rubbers, in which case tires and reinforced plastic tubing could be considered as examples. We could also envision more use being made of liquid-swollen soft structures. For purely mechanical systems, this may not make much sense, but in active systems, such as batteries or muscle-like actuators, a liquid component is necessary and should probably be viewed as a soft material rather than simply as a liquid to be contained.

B. Surfactants and Self-Assembly

Self-assembly is a hallmark of biological systems, including assembly of protein subunits into holoenzymes, of proteins and nucleic acids into virus particles, and of tropocollagen into collagen fibers. There has been increasing interest in synthetic self-assembly. In addition to the assembly of molecules with complementary hydrogen bonding to form supramolecular clusters, there are many papers on the assembly of charged polymers and particles to form multilayers and on the assembly of particles or particles and films coated with complementary biological recognition molecules, such as the biotin–streptavidin system.

Work at the Mobil Corporation showed that mesoporous silica could be formed by hydrolysis of tetraethoxysilane entrained in a highly concentrated water/silane/surfactant system. In this regime, the three-component mixture forms ordered structures with a range of symmetries. In one hexagonal phase, rods of water are surrounded by surfactant and embedded in a hydrophobic silane matrix. Hydrolysis of the silane under suitable conditions, followed by drying and sintering, results in a porous silica with aligned pores of a few nanometers' diameter.

The growth of the lyotropic liquid crystal precursor is very sensitive to the environment. Ozin and co-workers have shown that complex particle morphologies can result from growth of these mesoporous structures in quiescent solutions as diffusion fields and surface forces interact. Several workers have shown how the direction of the rods or plates of silica can be controlled. Polymers can be introduced to form composite structures that are very reminiscent of some biological composites. This does seem to parallel the proposed importance of liquid crystals in the growth of many biological structures.

While one would expect that this approach could be extended to many other material combinations, the rules are not understood. Efforts to form similar structures other than oxides, such as titania, or various crystalline materials, have been only partly successful. Possibly, any rapid or localized conversion process also disrupts the liquid crystalline organization. Stupp and co-workers (2000) have produced a range of amphiphiles that assemble into various ribbon and wedge structures, and the authors have explored their catalytic activity.

C. Inorganic Particle Formation

Coccolithophores, single-celled marine algae, assemble an external skeleton from single crystals of calcite with very complex shapes. The growing crystal is surrounded by a lipid membrane that controls the growth orientation in the crystal, but it is not known exactly how this is achieved. Sponges and diatoms show similar close control of the shape of silica particles on the micron scale. Magnetotatic bacteria form single-domain iron oxide (magnetite) crystals, with a very closely controlled size of a few nanometers, which then aggregate into magnetic chains (Fig. 1). These biological examples all involve growth within a compartment surrounded by a membrane. For sponge spicules there is an organic template on which the mineral grows. In other cases, there may be specific

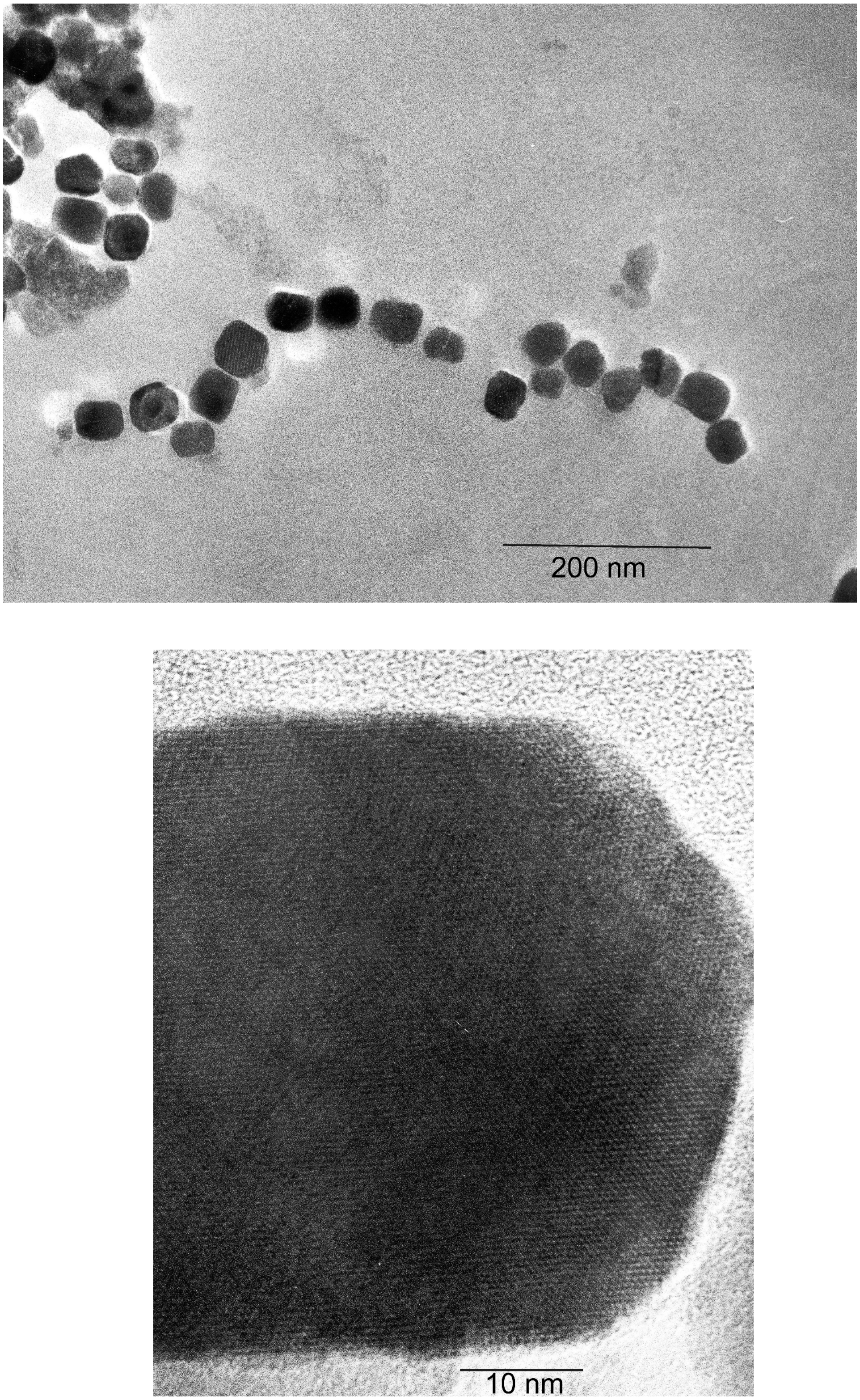


FIGURE 1 Magnetic particles from a magnetotactic bacterium, showing chain formation. [Courtesy of Prof. S. Seraphin, University of Arizona.]

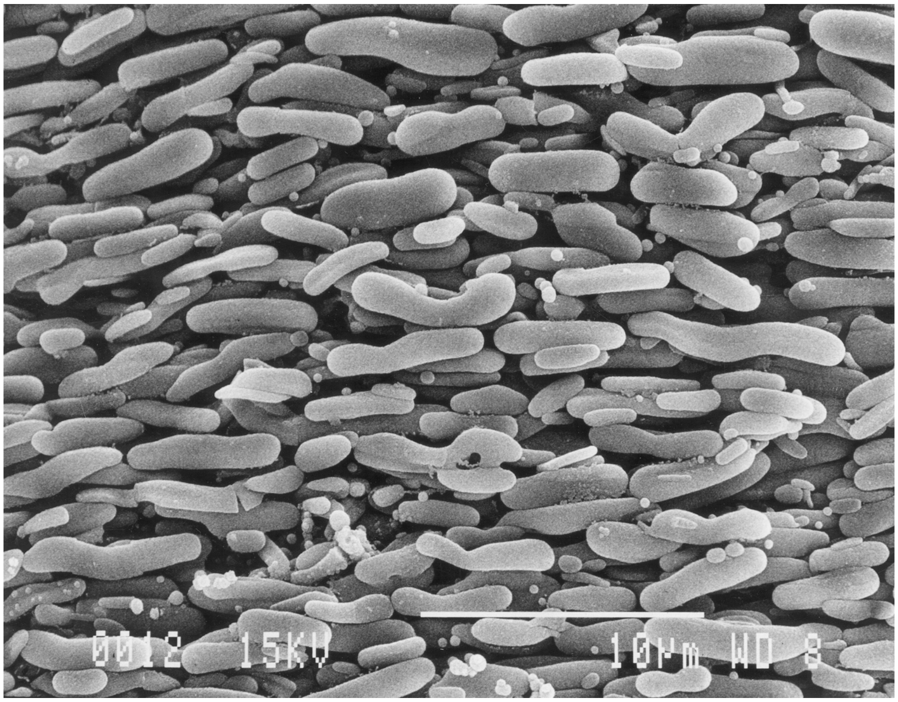


FIGURE 2 Elongated titania particles formed by impregnation of a stretched two-phase polymer with titanium alkoxide, followed by hydrolysis.

nucleation sites on the membrane surface. This suggests synthetic approaches where inorganic particles are grown within a micromold, a predefined space.

There have been many efforts to grow particles in liposomes, spherical shells with a lipid bilayer wall. One solution is trapped when the liposome is formed and precipitation occurs when a second reagent, often base, diffuses in through the wall. Generally, the trapped solution must be quite dilute in order to avoid destabilizing the liposome. As a result, the precipitate therefore only occupies a small part of the internal volume. A method is needed to introduce a continuous feed of both reagents through the membrane.

There have also been many efforts to grow particles in multiphase polymer systems, such as block copolymers. For instance, a two-phase polymer can be soaked in one reagent, which selectively absorbs into one phase. A cadmium salt could be taken up by a polyether phase. Subsequent treatment with hydrogen sulfide results in precipitation of cadmium sulfide within the polyether. While the volume fraction of sulfide formed is quite small, repeated cycles can give rise to higher fractions of the mineral. Work on precipitation in lyotropic liquid crystalline amphiphile solutions has led to composites structures.

One characteristic of biological minerals is their elaborate shapes. For mechanical reinforcement of soft matrices, a filler should be a high-aspect ratio rod or plate particles. These may grow as a natural outcome of differing crystal growth rates along different crystal axes. However, most simple minerals are not sufficiently anisotropic to form such elongated particles. The shape of biological crystals is probably controlled by selective inhibition of growth on specific crystal faces. Very elongated silica particles are also formed in lipid vesicles but we do not know how important these are for controlling shape in crystals. While synthetic methods offer many ways of controlling particle size, we lack good methods for controlling shape. One route is to form long micromolds by phase separation in a two-phase polymer, which is cold-drawn to elongate the included phase. The included phase is then swollen with a metal alkoxide, which is hydrolyzed to oxide (Fig. 2).

D. Tough-Layered Structures

Large organisms operate in an environment that subjects them to fluctuating forces, from the action of wind and water on plants and from the locomotion of animals. These fluctuating forces and collisions will often result in local damage that should not lead to catastrophic failure. As a result, many biological tissues contain structural features that add toughness without severely compromising stiffness or strength.

In mollusk shells and teeth, toughness may arise from added polymeric layers or from very fibrous structures, as discussed by Heuer. The addition of polymer could be especially effective if the polymer structure is capable

of a large extension breaking after yield, such as occurs in folded beta-sheet proteins, where the unfolding of the chains leads to a large energy absorption, as has been discussed by Morse (1999). However, polymer layers will also lead to a significant loss of elastic modulus when compared to a wholly inorganic material.

Much effort has gone into increasing the strength of ceramic materials by improving their toughness. There was initially much effort devoted to fibrous materials, particularly silicon nitride, but the need for reliable, inexpensive processing methods was not met. Attention then switched to ceramic composites with fibrous reinforcements that would add toughness. Mostly these proved to be unstable or reactive at the temperatures needed for turbine engines, while still being difficult to process. There have been many studies of ceramic/metal layered structures, where the metal layers take the role of the polymers in mollusk shell. Such materials are of interest both for use in engines and for armor. There is also interest in incorporating structured porosity in materials to improve fracture properties or to modify elastic properties such as Poisson's ratio or piezoelectric response.

In the case of metal/ceramic layered materials, the key problem is to absorb as much energy as possible within the metal layers. Hwu and Derby have shown that the major energy absorption is due to metal drawing across the gap after the crack has passed; however, the extent of deformation occurring during the drawing is limited by the thinness of the metal layers. As a result, substantial increases in toughness were only seen when the volume fraction of metal was quite large. In contrast, mollusk shells show significant toughness with only a small percent of polymer. It is likely that some combination of embedded metal, porosity, and interfacial debonding can give much enhanced toughness or impact resistance to ceramics, but current methods for modeling fracture do not provide enough guidance for these designs and the biological models are only just being interpreted.

Studies on the influence of polymers on the crystallization of calcium carbonates *in vitro* have shown that metastable liquid complexes with anionic polymers can be formed that subsequently transform to calcite. Studies of carbonate biomineralization have shown that amorphous calcium carbonate forms either transiently or as a stable phase. It has also been shown that absorbed proteins modify the fracture properties of carbonates. In particular, fracture surfaces show smooth conchoidal fracture, like glass, rather than the faceted fracture characteristic of normal crystals. This suggests that we have much to learn about the modification of the properties of crystals and amorphous solids by entrained polymer (see Fig. 3). There are parallels with the Lanxide process for

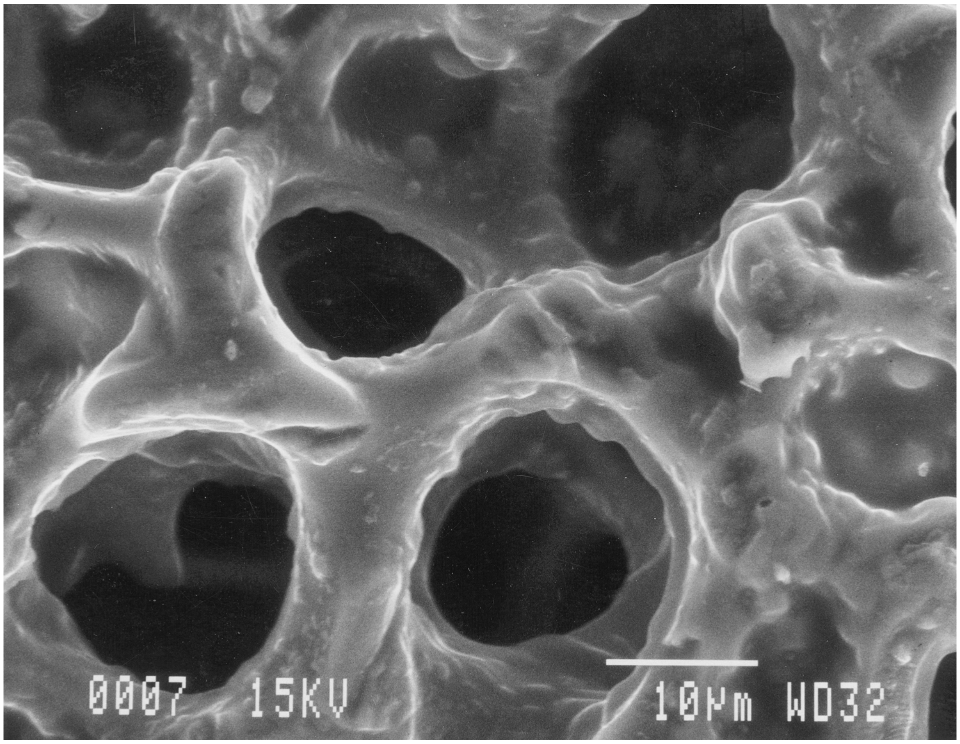


FIGURE 3 Porous interior of a sea urchin spine; calcite single crystal. Polymer coating may serve to prevent fracture from surface damage; entrained polymer may enhance the toughness.

toughening alumina ceramic by entrained aluminum metal particles.

E. Biomimetic Processing Methods

Freeform fabrication methods allow objects to be built as a series of layers directly from a three-dimensional computer representation. These allow only one material to be used at a time, or a material plus a soluble support structure. It is clearly feasible to combine several different materials into a single object, which would allow the building of something much closer to an organism. A report on simple robots, which were allowed to evolve in a virtual environment and then were built by freeform fabrication, shows how evolutionary methods might be applied to manufactured objects.

To build a crude organism would require resolution at the scale of about 10 μm. Most current freeforming methods allow resolution down to about 100 μm and a limited materials set. Microcontact printing and related methods allow much higher resolution, down to about 1 μm, but effectively are restricted to one layer. There is much current interest in ink-jet printing methods, which could provide the required 10-μm resolution while allowing many layers to be deposited.

In both synthetic and biological structures, it is useful to keep in mind a distinction between design and patterning. Phase separation, crystallization, and aggregation processes can give rise to patterns in two and three dimensions on a scale from millimeters to nanometers which reflect the kinetics of the separation and diffusion processes. To form working devices or organisms, we need to build to a nonrecurring design, which may include patterned elements. If we seek to adopt biomimetic processes, we will need to exploit self-assembling structures and patterns but within an overall design.

In current silicon technology, photolithographic methods can form two-dimensional designs down to less than 1 μm. Much finer resolutions are achievable in laboratory methods. Three-dimensional designs can be formed to below 1 μm using two-photon methods but commercial freeforming methods are limited to about 100 μm. In biology, there are many examples of patterns forming at the nanometer level but most designs are on the scale of individual cells, a few tens of microns. There are structures, such as sensing hairs, which are much finer but the spacing between them is still on the 10-μm scale.

F. Cell Adhesion and Tissue Engineering

Biomedical engineering is becoming more concerned with the problems of the long-term biocompatibility of synthetic implants. Material wear and degradation and tissue loss or chronic inflammation due to changing mechanical loads will limit the lifetime of most implants to about 10 years. In addition, biocompatibility is not a material property but is very dependent on the specific environment of the implant, in terms of implant site in the body, animal species, and animal age. Strategies to eliminate these problems include the use of biodegradable materials, which will eventually be replaced by natural tissue, and tissue engineered implants, where as much of the device as possible is formed from tissue grown *in vitro* on a synthetic support before implantation.

Tissue engineering imposes a need to understand the interactions between neighboring cells, between cells, and between cells and synthetic surfaces. Cell binding to a biological surface proceeds through a series of stages including physical adsorption, interaction between surface macromolecules and specific binding sites on the surface, reorganization of the macromolecules in the cell membrane to bring more binding sites into contact with the surface, and then specific changes in the cell induced by the surface.

At the physical adsorption level, studies of cell attachment to self-assembled monolayers on silicon or glass have shown that attached polyethyleneoxide chains form a structureless hydrophilic barrier layer and will prevent adsorption. Different polar end groups will allow cells to absorb, and printed surface patterns with binding and nonbinding areas can be used to control cell shape. Cells recognize and bind to simple short sequences of amino acids in a protein exposed on a surface. RGD (arginine–glycine–aspartic acid) and RADS (arginine–alanine–aspartic acid–serine) are sequences that can be used to induce strong cell attachment. Thus, suitable polymers can be produced which will promote formation of particular cell types to the surfaces for tissue engineering.

In the case of bone implants, there has been much interest in the use of bone morphogenic proteins, now produced from cloned bacteria. One group of a family of cell-signaling proteins are known to induce the development of bone or other tissues when locally released in the body. A similar set of signals induces growth of small blood vessels and is an important factor in the development of many cancers. By allowing implants to release such signaling proteins, it should be possible to speed the integration of the implant; however, much remains to be learnted about the appropriate concentrations, gradients, and timing of these signals. All of these efforts take us in the direction of making synthetic organs that look more like a natural organ transplanted from an identical twin.

Belcher and co-workers have recently demonstrated that a phage library, displaying 10^9 different peptide sequences at the surface, can be used to identify short peptide sequences that selectively bind to inorganic semiconductor surfaces. Sarikaya and co-workers have developed

a similar method for control of gold precipitation using *Escherichia coli* genetics.

VIII. APPLICATIONS OF BIOMIMETIC MATERIALS

Vogel (1998) has discussed the importance of biomimesis in engineering, asking whether it is possible to unequivocally attribute any engineering advance to a biological inspiration. There are no unambiguous examples, and there are cases, such as the attempts to imitate flapping flight, where a biological analogy may have inhibited progress. At the same time, many engineering advances clearly drew some inspiration from biology. In chemistry and materials, studies of the structure and properties of biological materials do suggest alternative approaches to particular problems and illustrate new properties that should be achievable in synthetic materials.

Biomimesis clearly will have an important role in new biomedical devices and in new devices that try to combine biological structures or organisms with electronics in sensors and actuators. In structural materials, the obvious place for advance is the introduction of more toughness into synthetic composites and ceramics. In electronics there are not yet many signs of a conjunction between the hard, high-resolution, two-dimensional world of silicon and the soft, larger scale, three-dimensional design of the brain and nervous system.

SEE ALSO THE FOLLOWING ARTICLES

Biomaterials, Synthesis, Fabrication, and Applications • Biopolymers • Glycoconjugates and Carbohydrates • Materials Chemistry • Separation and Purification of Biochemicals • Tissue Engineering

BIBLIOGRAPHY

Baer, E., Cassidy, J. J., and Hiltner, A. (1991). "Hierarchical structure of collagen composite systems: lessons from biology," *Pure Appl. Chem.* **63,** 961–973.

Baskaran, S., Nunn, S. D., Popovic, D., and Halloran, J. W. (1993). "Fibrous monolithic ceramics," *J. Amer. Ceram. Soc.* **76,** 2209–2224.

Beniash, E., Addadi, L., and Weiner, S. (1999). "Cellular control over spicule formation in sea urchin embryos: a structural approach," *J. Struct. Biol.* **125,** 50–62.

Bertrand, P., Jonas, A., Laschewsky, A., and Legras, R. (2000). "Ultrathin polymer coatings by complexation of polyelectrolytes at interfaces: suitable materials, structure and properties," *Macromol. Rapid Commun.* **21,** 319–348.

Brown, S., Sarikaya, M., and Johnson, E. (2000). "A genetic analysis of crystal growth," *J. Molec. Biol.* **299,** 725–735.

Burdon, J., Oner, M., and Calvert, P. (1996). "Growth of oxalate crystals on films of acrylate polymers," *Mater. Sci. Eng. C* **4**.

Calvert, P. (1996). "Biomimetic processing." *In* "Materials Science & Technology," vol. 17B, (R. J. Brook, ed.), pp. 51–82, VCH Publishers; Weinheim.

Calvert, P., and Crockett, R. (1997). "Chemical solid free-form fabrication: making shapes without molds," *Chem. Mater.* **9,** 650–663.

Calvert, P., and Rieke., P. (1996). "Biomimetic mineralization in and on polymers," *Chem. Mater.* **8,** 1715–1727.

Calvert, P. D. (1994). "Polymers for New Materials," *Polymer* **35,** 4484–4488.

Cha, J., Stucky, G., Morse, D., and Deming, T. (2000). "Biomimetic synthesis of ordered silica structures mediated by block copolypeptides," *Nature* **403,** 289–292.

Colgin, M. A., and Lewis, R. V. (1998). "Spider minor ampullate silk proteins contain new repetitive sequences and highly conserved non-silk-like 'spacer regions,'" *Protein Sci.* **7,** 667–672.

Currey, J. D., Zioupos, P., and Sedman, A. (1995). "Microstructure-Property relationships in vertebrate bony hard tissues." *In* "Biomimetics" (M. Sarikaya and I. A. Aksay, eds.), pp. 117–144, AIP Press, Woodbury, NY.

De Guire, M., Niesen, T., Supothina, S., Wolff, J., Bill, J., Sukenik, C., Aldinger, F., Heuer, A., and Ruhle, M. (1998). "Synthesis of oxide and non-oxide inorganic materials at organic surfaces," *Z. Metallk.* **89,** 758–766.

Fan, H., Lu, Y., Stump, A., Reed, S., Baer, T., Schunk, R., Perez-Luna, V., Lopez, G., and Brinker, C. (2000). "Rapid prototyping of patterned functional nanostructures," *Nature* **405,** 56–60.

Frankel, R., Bazylinski, D., and Schuler, D. (1998). "Biomineralization of magnetic iron minerals in bacteria," *Supramolecular Sci.* **5,** 383–390.

Fritz, M., Belcher, A. M., Radmacher, M., Walters, D. A., Hansma, P. K., Stuckey, G. D., Morse, D. E., and Mann, S. (1994). "Flat pearls from biofabrication of organized composites on inorganic substrates," *Nature* **371,** 49–51.

GiraudGuille, M. (1996). "Twisted liquid crystalline supramolecular arrangements in morphogenesis," *Int. Rev. Cytol. Surv. Cell Biol.* **166,** 59–101.

Gunderson, S. L., and Schiavone, R. C. (1995). "Microstructure of an insect cuticle and applications to advanced composites." *In* "Biomimetics" (M. Sarikaya and I. A. Aksay, eds.), pp. 163–198, AIP Press, Woodbury, NY.

Hayashi, C. Y., and Lewis, R. V. (2000). "Molecular architecture and evolution of a modular spider silk protein gene," *Science* **287,** 1477–1479.

Iijima, M., and Moriwaki, Y. (1999). "Effects of ionic inflow and organic matrix on crystal growth of octacalcium phosphate relevant to tooth enamel formation," *J. Crystal Growth* **199,** 670–676.

Kamat, S., Su., X., Ballarini, R., and Heuer, A. H. (2000). "Structural basis for the fracture toughness of the shell of the conch *Strombus gigas*," *Nature* **405,** 1036–1040.

Kane, R. S., Cohen, R. E., and Silbey, R. (1996). "Synthesis of PbS nanoclusters within block copolymer nanoreactors," *Chem. Mater.* **8,** 1919–1924.

Kaplan, D. L., Mello, C. M., Arcidiacono, S., Fossey, S., Senecal, K., and Muller, W. (1997). "Silk." *In* "Protein Based Materials" (K. P. McGrath and D. L. Kaplan, eds.), pp. 103–132, Birkhauser, Boston.

Kasapi, M. A., and Gosline, J. M. (1997). "Design complexity and fracture control in the equine hoof wall," *J. Exp. Biol.* **200,** 1639–1659.

Kroger, N., Deutzmann, R., and Sumper, M. (1999). "Polycationic peptides from diatom biosilica that direct silica nanosphere formation," *Science* **286,** 1129–1132.

Lehn, J. M. (1990). "Perspectives in supramolecular chemistry—from molecular recognition towards molecular information-processing and self-organization," *Angew. Chem. Int. Ed.* **29,** 1304–1319.

Levi, C., Barton, J. L., Guillemet, C., Lebras, E., and Lehuede, P. (1989). "A remarkably strong natural glassy rod: the anchoring spicule of *Monoraphis* sponge," *J. Mater. Sci. Lett.* **8,** 337–339.

Li, C.-W., and Volcani, B. E. (1984). "Silicification in diatom walls," *Philos. Trans. R. Soc. London* **B304,** 519–528.

Lipson, H., and Pollack, J. B. (2000). "Automatic design and manufacture of robotic lifeforms," *Nature* **406,** 974–978.

Lowenstam, H. A., and Weiner, S. (1989). "On Biomineralization," Oxford University Press, London.

Mann, S., Burkett, S., Davis, S., Fowler, C., Mendelson, N., Sims, S., Walsh, D., and Whilton, N. (1997). "Sol-gel synthesis of organized matter," *Chem. Mater.* **9,** 2300–2310.

Miyaji, F., Kim, H., Handa, S., Kokubo, T., and Nakamura, T. (1999). "Bonelike apatite coating on organic polymers: novel nucleation process using sodium silicate solution," *Biomaterials* **20,** 913–919.

Morse, D. (1999). "Silicon biotechnology: harnessing biological silica production to construct new materials," *Trends Biotechnol.* **17,** 230–232.

National Materials Advisory Board (1994). "Hierarchical Structures in Biology as a Guide for New Materials Technology," National Academy Press, Washington, D.C.

Privman, V., Goia, D., Park, J., and Matijevic, E. (1999). "Mechanism of formation of monodispersed colloids by aggregation of nanosize precursors," *J. Colloid Interface Sci.* **213,** 36–45.

Rajam, S., Heywood, B. R., Walker, J. B. A., Mann, S., Davey, R. J., and Birchall, J. D. (1991). "Oriented crystallization of $CaCO_3$ under compressed monolayers. Part 1. Morphological studies of mature crystals," *J. Chem. Soc., Faraday Trans.* **87,** 727–734.

Raz, S., Weiner, S., and Addadi, L. (2000). "Formation of high-magnesian calcites via an amorphous precursor phase: possible biological implications," *Adv. Mater.* **12,** 38–40.

Schaffer, T. E., Ionescu-Zanetti, C., Proksch, R., Fritz, M., Walters, D. A., Almqvist, N., Zaremba, C. M., Belcher, A. M., Smith, B. L., Stucky, G. D., Morse, D. E., and Hansma, P. K. (1997). "Does abalone nacre form by heteroepitaxial nucleation or by growth through mineral bridges?" *Chem. Mater.* **9,** 1731–1740.

Sellinger, A., Weiss, P., Nguyen, A., Lu, Y., Assink, R., Gong, W., and Brinker, C. (1998). "Continuous self-assembly of organic-inorganic nanocomposite coatings that mimic nacre," *Nature* **394,** 256–260.

Sigmund, O., Torquato, S., and Aksay, I. A. (1998). "On the design of 1–3 piezocomposites using topology optimization," *J. Mater. Res.* **13,** 1038–1048.

Simpson, T., and Volcani, B. (1981). "Silicon and Siliceous Structures in Biological Systems," Springer–Verlag, Berlin.

Stupp, S., Pralle, M., Tew, G., Li, L., Sayar, M., and Zubarev, E. (2000). "Self-assembly of organic nano-objects into functional materials," *Mater. Res. Soc. Bull.* **25**(4), 42–48.

Vincent, J. F. V. (1980). "Insect cuticle: a paradigm for natural composites." *In* "The Mechanical Properties of Biological Materials" (J. F. V. Vincent and J. D. Currey, eds.), pp. 183–210, Cambridge University Press, Cambridge, U.K.

Vogel, S. (1998). "Cats' Paws and Catapults," W. W. Norton, New York.

Wainwright, S. A., Biggs, W. D., Currey, J. D., and Gosline, J. M. (1986). "Mechanical Design in Organisms," Princeton University Press, Princeton, NJ.

Weiner, S., Addadi, L., and Wagner, H. D. (2000). "Materials design in biology," *Mater. Sci. Eng. C* **11,** 1–8.

Whaley, S. R., English, D. S., Hu, E. L., Barbara, P. F., and Belcher, A. M. (2000). "Selection of peptides with semiconductor binding specificity for directed nanocrystal assembly," *Nature* **405,** 665–668.

Xia, Y., and Whitesides, G. (1998). "Soft lithography," *Ann. Rev. Mater. Sci.* **28,** 153–184.

Yang, S., Yang, H., Coombs, N., Sokolov, I., Kresge, C., and Ozin, G. (1999). "Morphokinetics: growth of mesoporous silica curved shapes," *Adv. Mater.* **11,** 52–55.

Bioreactors

Yusuf Chisti
University of Almería

Murray Moo-Young
University of Waterloo

I. Introduction
II. Bioreactor Systems
III. Considerations for Bioreactor Design
IV. Concluding Remarks

GLOSSARY

Austenitic stainless steels Stainless steels with an specific type of nonmagnetic crystal structure.

Downcomer The region of a an airlift bioreactor where the gas–liquid dispersion flows downwards.

Enzymes Protein molecules that catalyze the various biochemical reactions.

Heterotrophic growth Growth in which the carbon and energy for building the biomass are derived exclusively from organic chemicals.

Mass transfer Molecular-level transport of any substance through any medium.

Organelles Well-defined structures associated with specific functions in a cell.

Photomixotrophic culture A culture that obtains a part of the carbon needed for making the biomass from an organic chemical and another part from carbon dioxide via photosynthesis

Phototrophic growth Growth in which the carbon needed to make the biomass comes from fixation of carbon dioxide via photosynthesis.

Pressure drop Loss of pressure with distance downstream from any point in tubes, channels, and other flow devices.

Product (or substrate) inhibition A situation in which the increasing concentration of the product (or substrate) of a reaction slows down the rate of the reaction by interfering with the enzyme(s) that catalyze the reaction.

Protoplast A cell with its wall removed.

Reduced substrate A substrate that contains relatively little oxygen within its molecules.

Riser The region of an airlift bioreactor where the gas-liquid dispersion flows upwards.

Solidity ratio The ratio of the swept area to the total projected area of the impeller blades, as viewed from directly overhead an installed impeller.

Substrate Any compound that is modified by a cell or an enzyme.

Water-for-injection (WFI) Highly purified water that conforms to the WFI specifications of the United States Pharmacopoeia.

A BIOREACTOR is any device or vessel that is used to carry out one or more biochemical reactions to convert any starting material (or raw material or *substrate*) into some product. The conversion occurs through the action of a *biocatalyst*—enzymes, microorganisms, cells of animals and plants, or subcellular structures such as chloroplasts and mitochondria. The starting substrate may be a simple

organic chemical (e.g., sugar and penicillin), an inorganic chemical such as carbon dioxide, or a poorly defined complex material such as meat and animal manure. The product of the conversion may be cells (or biomass), viruses, and chemicals of various kinds. The processes occurring in a bioreactor may be summarized as follows:

$$\text{Substrates} \xrightarrow{\text{biocatalyst(s)}} \text{products.}$$

Many different kinds of bioreactors are available and sometimes a given type may be operated in different ways to obtain different results. The different bioreactor designs are needed to accommodate the great diversity of substrates, products, and biocatalysts and the different requirements of the different bioconversion processes.

I. INTRODUCTION

Bioreactors are used in all kinds of bioprocesses, including those for making foods such as soy sauce; those for treating domestic and industrial wastewater; and ones for making vaccines, antibiotics, and many other useful chemicals. Bioreactors that produce microbial cells and cells of animals and plants, are known as *fermenters*. In addition to cells, a fermenter may also produce other chemicals, or convert (or biotransform) a chemical added to the fermenter, to a different molecule. A fermenter may contain either a single cell type (i.e., *monoseptic* operation), or a mixed population of different kinds of cells. Fermenters that operate monseptically are designed as sealed units, with barriers that prevent ingress of contaminating microorganisms from the environment. Other types of bioreactors may contain only nonviable entities (i.e., ones that cannot multiply) including cells, isolated enzymes, and organelles obtained from cells. The cells and organelles may be freely suspended in an aqueous medium, or they may be confined by various methods of immobilization. Unlike cells and organelles, enzymes are usually soluble in aqueous media; for repeated use, a soluble enzyme may be retained in the bioreactor by ultrafiltration membranes, or the enzyme may be immobilized in an insoluble matrix.

II. BIOREACTOR SYSTEMS

In most bioprocessing situations cells and biocatalysts are submerged and suspended in a broth that sustains live cultures and dissolves the chemicals that are being modified by the action of the biocatalyst. Bioreactors for submerged processing are generally quite different from those used in solid-state cultivation. Solid-state fermentations are carried out with a moistened solid substrate in the absence of free water, e.g., during composting, making of hard cheeses, and fermentation of cocoa beans for chocolate. Submerged processing is widely used in treatment of wastewater and production of vaccines, antibiotics, and many other useful products. Bioreactors for submerged and solid-state processes are discussed next.

A. Submerged Culture

1. Mechanically Stirred Tank Bioreactors

Stirred tank bioreactors consist of a cylindrical vessel with a motor driven central shaft that supports one or more agitators (Fig. 1). Different kinds of agitators are used in different applications. Microbial culture vessels are generally provided with four baffles placed equidistant around the periphery of the tank. The baffles project into the vessel from near the walls (Fig. 1). The baffles run the entire working height of the vessel and they prevent swirling and vortexing of the fluid. The baffle width is 1/10 or 1/12 of the tank diameter. A gap of about 1.5% of tank diameter is left between the wall and the baffle to prevent stagnation of fluid near the wall. The working aspect ratio of the vessel is between 3 and 5, except in animal cell culture applications where aspect ratios do not normally exceed 2. Often, the animal cell culture vessels are unbaffled.

The number of impellers used depends on the aspect ratio of the vessel. The lowermost impeller is located about one third of the tank diameter above the bottom of the tank. Additional impellers are spaced with

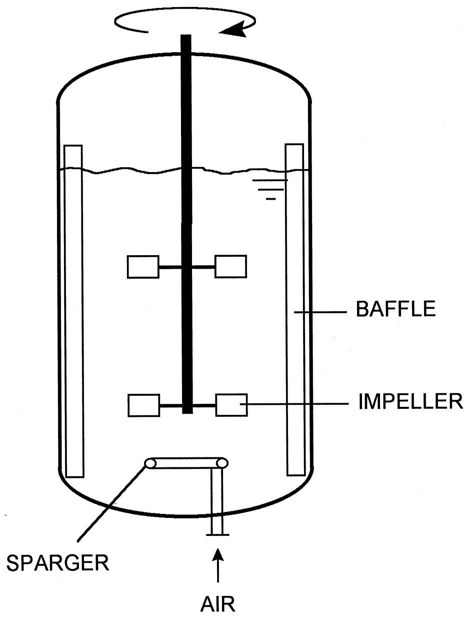


FIGURE 1 Mechanically stirred bioreactor.

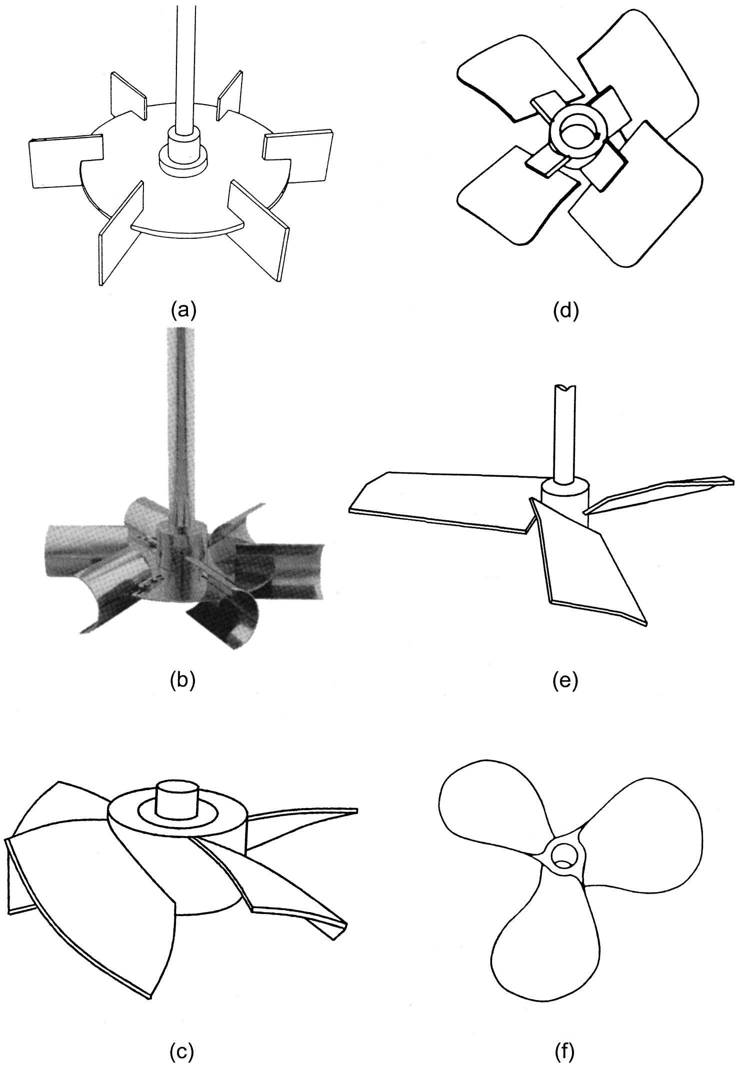


FIGURE 2 Agitators or impellers used in stirred bioreactors: (a) Rushton disc turbine; (b) concave bladed turbine; (c) Prochem Maxflo T hydrofoil; (d) Lightnin' A315 hydrofoil; (e) Chemineer hydrofoil; and (f) marine propeller.

1.2-impeller-diameter distance in between. The impeller diameter is about 1/3 of the vessel diameter for gas dispersion impellers such as Rushton disc turbines and concave bladed impellers (Fig. 2a, b). Larger hydrofoil impellers (Fig. 2c–e) with diameters of 0.5 to 0.6 times the tank diameter are especially effective bulk mixers and are used increasingly. High solidity ratio hydrofoils (Fig. 2c, d) are good for highly viscous mycelial broths. Animal cell culture vessels typically employ a single large-diameter, low-shear impeller such as a marine propeller (Fig. 2f). Oxygen is provided typically by sparging the broth with sterile air. In microbial fermenters, gas is sparged below the lowermost impeller using a perforated pipe ring sparger with a ring diameter that is slightly smaller than that of the impeller (Fig. 1). A single hole sparger discharging the gas below the impeller at the tank centerline is used sometimes. Aeration velocity is usually kept at less than 0.05 ms^{-1}, or the mixing effectiveness of the impeller will be reduced. In bioreactors for animal cell culture, the aeration velocities are lower, usually less than 0.01 m s^{-1}, and the gas is sparged such that it does not rise through the region swept by the impeller. Mixing, oxygen transfer, and heat transfer improve with increasing agitation and aeration rates. In low viscosity media, the type of impeller has little effect on the gas–liquid mass transfer rate, so long as the power input per unit liquid volume is kept unchanged.

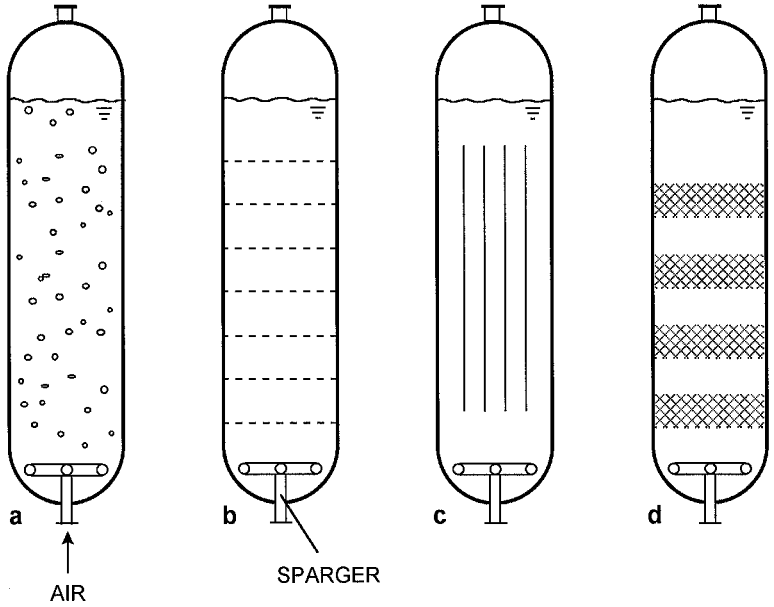


FIGURE 3 Bubble columns: (a) the basic design, (b) a column with transverse perforated baffle plates, (c) a column with vertical baffles, and (d) a column with corrugated sheet static mixers for gas dispersion.

2. Bubble Columns

A bubble column consists of a gas sparged pool of liquid or slurry (Fig. 3a). Usually, the column is cylindrical and the aspect ratio is between 4 and 6. This basic design may be modified by placing various kinds of internals—e.g., horizontal perforated plates, vertical baffles, and corrugated sheet packings—inside the vessel (Fig. 3b–d). Gas is sparged at the base of the column through perforated pipes, perforated plates, or sintered glass or metal microporous spargers. Oxygen transfer, mixing, and other performance factors are influenced mainly by the gas flow rate and the properties of the fluid. The column diameter does not affect its behavior so long as the diameter exceeds 0.1 m. One exception is the mixing performance. For a given gas flow rate, the mixing improves with increasing vessel diameter. Mass and heat transfer performance improves as gas flow rate is increased. In bubble columns the maximum aeration velocity usually remains less than 0.1 ms^{-1}. The liquid flow rate does not influence the gas–liquid mass transfer coefficient so long as the superficial liquid velocity remains below 0.1 ms^{-1}.

Bubble columns with recirculation and airlift bioreactors are especially suited to hairy root culture of plant cells. The rootlets tend to grow as an entangled static mass with a doubling time of about 2 days. The fluid flowing past the roots supplies oxygen and other nutrients. A bubble column bioreactor with the hairy root mass confined between two perforated retention plates, is shown in Fig. 4. The nutrient medium flowing into the column may be oxygenated in a separate column, or air may be bubbled in the column that contains the root mass. External oxygenation is suitable when the conditions (oxygen consumption, fluid residence time) in the column with the root mass are such that the spent medium leaving the column is not totally depleted of oxygen.

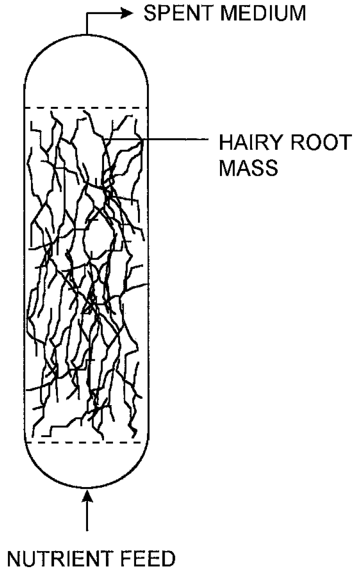


FIGURE 4 Bubble column bioreactor for hairy root cultivation.

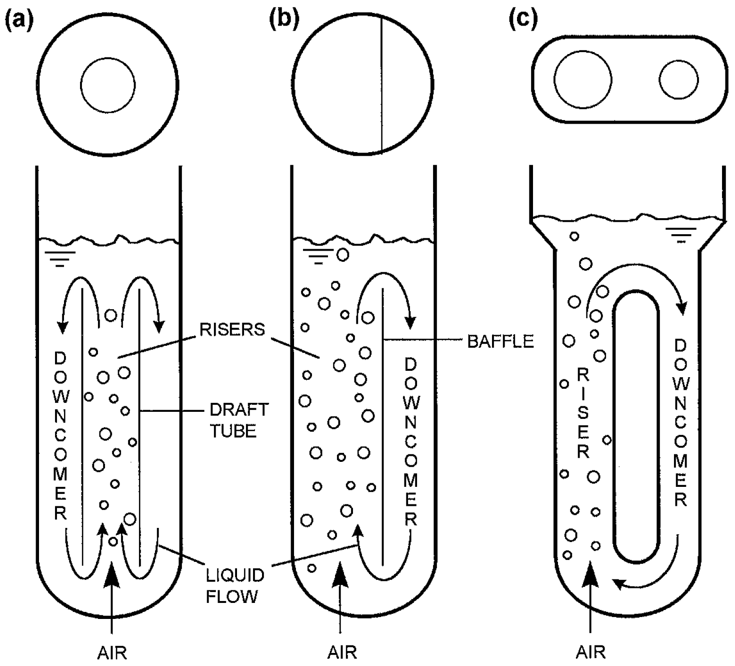


FIGURE 5 Airlift bioreactors: (a) draft-tube internal-loop configuration, (b) a split-cylinder device, and (c) an external-loop system.

3. Airlift Bioreactors

In airlift bioreactors the fluid volume of the vessel is divided into two interconnected zones by means of a baffle or draft-tube (Fig. 5). Only one of these zones is sparged with air or other gas. The sparged zone is known as the riser; the zone that receives no gas is the downcomer (Fig. 5a–c). The bulk density of the gas-liquid dispersion in the gas-sparged riser tends to be less than the bulk density in the downcomer; consequently, the dispersion flows up in the riser zone and downflow occurs in the downcomer. Sometimes the riser and the downcomer are two separate vertical pipes that are interconnected at the top and the bottom to form an external circulation loop (Fig. 5c). External-loop airlift reactors are less common in commercial processes compared to the internal-loop designs (Fig. 5a, b). The internal-loop configuration may be either a concentric draft-tube device or an split-cylinder (Fig. 5a, b). Airlift reactors have been successfully employed in nearly every kind of bioprocess—bacterial and yeast culture, fermentations of mycelial fungi, animal and plant cell culture, immobilized enzyme and cell biocatalysis, culture of microalgae, and wastewater treatment.

Airlift bioreactors are highly energy efficient relative to stirred fermenters, yet the productivities of both types are comparable. Heat and mass transfer capabilities of airlift reactors are at least as good as those of other systems, and airlift reactors are more effective in suspending solids than are bubble columns. For optimal gas–liquid mass transfer performance, the riser-to-downcomer cross-sectional area ratio should be between 1.8 and 4.3 in an airlift reactor. All performance characteristics of airlift bioreactors are linked ultimately to the gas injection rate and the resulting rate of liquid circulation. The liquid circulation velocity depends on the difference in gas holdup (i.e., the volume fraction of gas in the gas-liquid dispersion) between the riser and the downcomer. Liquid velocity is affected also by the geometry of the reactor and the viscosity of the fluid. In general, the rate of liquid circulation increases with the square root of the height of the airlift device. Consequently, the reactors are designed with high aspect ratios of at least 6 or 7, or even in the hundreds. Because circulation is driven by the gas holdup difference between the riser and the downcomer, circulation is enhanced if there is little or no gas in the downcomer. All the gas in the downcomer comes from being dragged in with the liquid as it flows into the downcomer from the riser near the top of the reactor (Fig. 5). Various designs of gas-liquid separators (Fig. 6) are sometimes used in the head zone to reduce or eliminate the gas carry over to the downcomer. Most gas–liquid separators work in one of two ways: either the horizontal flow path between the riser and the downcomer is extended (Fig. 6a) so that the liquid resides for a longer period in the head zone and this provides sufficient time for the gas bubbles to disengage; or the entrance region of the downcomer is expanded in cross section (Fig. 6b) so that the downward flow velocity of the liquid is reduced and it no longer drags gas bubbles into the downcomer. Relative

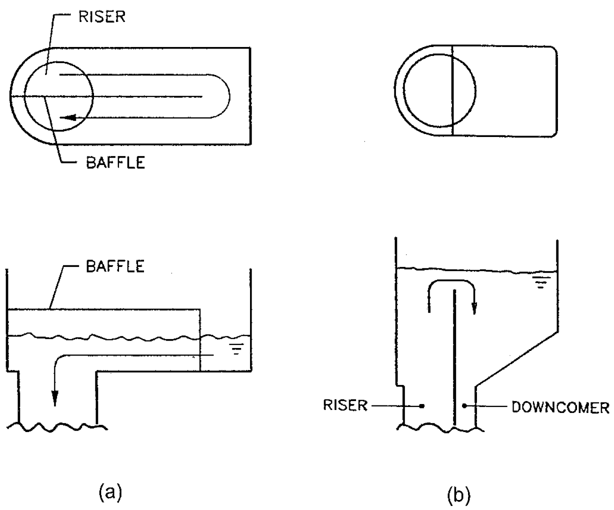


FIGURE 6 Gas–liquid separators for airlift bioreactors: (a) extended length of the flow path in the head zone, (b) enlarged entrance cross section of the downcomer zone.

to a reactor without a gas–liquid separator, installation of a suitably designed separator will always enhance liquid circulation, i.e., the increased driving force for circulation will more than compensate for any additional resistance to flow due to the separator.

4. Fluidized Beds

Fluidized bed bioreactors are suited to reactions involving a fluid-suspended particulate biocatalyst such as the immobilized enzyme and cell particles or microbial flocs. An up-flowing stream of liquid is used to suspend or "fluidize" the relatively dense solids (Fig. 7). Geometrically, the reactor is similar to a bubble column except that the cross section is expanded near the top to reduce the superficial velocity of the fluidizing liquid to a value below that needed to keep the solids in suspension (Fig. 7). Consequently, the solids sediment in the expanded zone and drop back into the narrower reactor column below; hence, the solids are retained in the reactor whereas the liquid flows out. A liquid fluidized bed may be sparged with air or some other gas to produce a gas–liquid–solid fluid bed. If the solid particles are too light, they may have to be artificially weighted, for example, by embedding stainless steel balls in an otherwise light solid matrix. A high density of solids improves solid–liquid mass transfer by increasing the relative velocity between the phases. Denser solids are also easier to sediment, but the density should not be too great relative to that of the liquid, or fluidization will be difficult.

Liquid fluidized beds tend to be fairly quiescent but introduction of a gas substantially enhances turbulence and agitation. Even with relatively light particles, the superficial liquid velocity needed to suspend the solids may be so high that the liquid leaves the reactor much too quickly, i.e., the solid–liquid contact time may be insufficient for the reaction and the liquid may have to be recycled to obtain a sufficiently long cumulative contact time with the biocatalyst. The minimum fluidization velocity—i.e., the superficial liquid velocity needed to just suspend the solids from a settled state—depends on several factors, including the density difference between the phases, the shape and diameter of the particles, and the viscosity of the liquid.

5. Packed Bed Bioreactors

A bed of solid particles usually with confining walls (Fig. 8) constitutes a packed bed. The biocatalyst is supported on or within the solid matrix that may be porous or a homogeneous non-porous gel. The solids may be ridged, or only slightly compressible. The particles may be randomly shaped (e.g., wood chips and rocks) or they may be uniform spheres, cylinders, cubes, or some other shape. A fluid containing dissolved nutrients and substrates flows through the solid bed to provide the needs of the immobilized biocatalyst. Metabolites and products are released into the fluid and are taken out with the flow. The flow may be upward or downward, but downflow under gravity (i.e., trickle bed operation) is the norm specially if the immobilized biocatalyst requires oxygen (Fig. 8). If the fluid flows up the bed, the maximum flow velocity is limited because the velocity cannot exceed the minimum

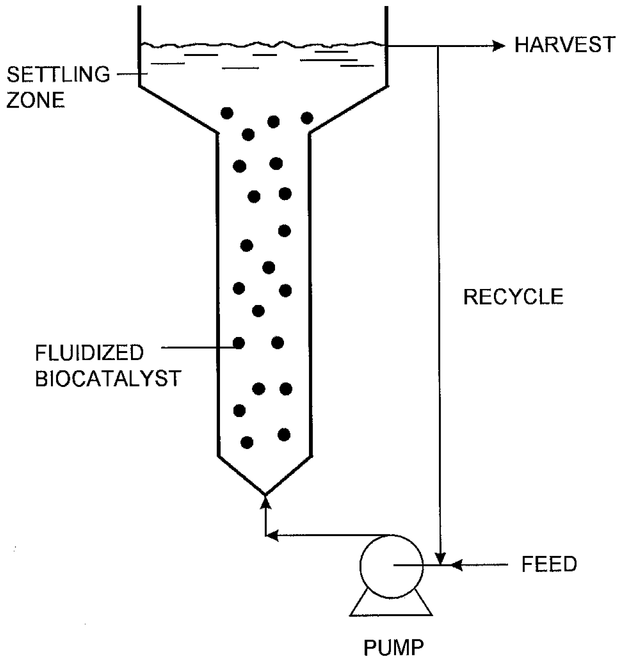


FIGURE 7 A fluidized bed bioreactor with recycle of medium.

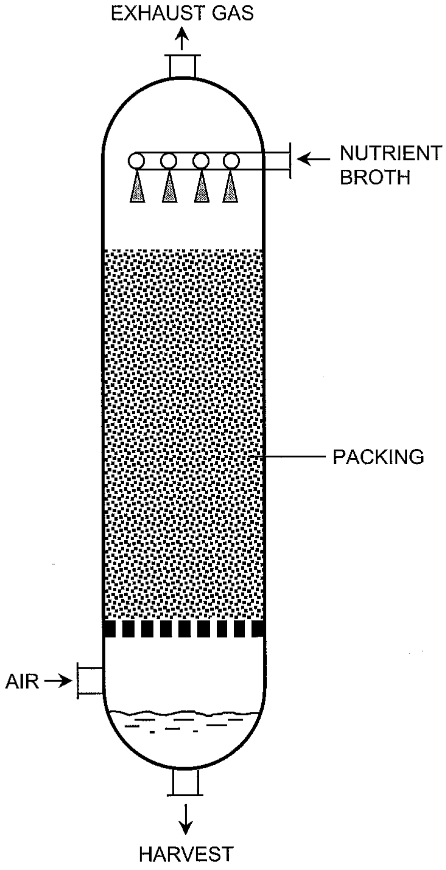


FIGURE 8 A packed bed bioreactor.

fluidization velocity or the bed will fluidize. The depth of the bed is limited by several factors, including the density and the compressibility of the solids, the need to maintain a certain minimal level of a critical nutrient such as oxygen through the entire depth, and considerations of the maximum acceptable pressure drop. For a given voidage—or solids-free volume fraction of the bed—the gravity driven flow rate through the bed declines if the depth of the bed is increased. Nutrients and substrates are depleted as the fluid moves down the bed. Conversely, concentrations of metabolites and products increase. Thus, the environment of a packed bed is nonhomogeneous, but concentration variations along the depth can be reduced by increasing the flow rate. Gradients of pH may occur if the reaction consumes or produces the H^+ ion. Because of poor mixing, pH control by addition of acid and alkali is nearly impossible. Beds with greater voidage permit greater flow velocities through them, but the concentration of the biocatalyst in a given bed volume declines as the voidage is increased. If the packing—i.e., the biocatalyst-supporting solids—is compressible, its weight may compresses the bed unless the packing height is kept low. Flow is difficult through a compressed bed because of a reduced voidage. Packed beds are used especially commonly as immobilized enzyme reactors and "biofilters" for the treatment of gaseous pollutants. Such reactors are particularly attractive for product inhibited reactions: the product concentration varies from a low value at the inlet of the bed to a high value at the exit; thus, only a part of the biocatalyst is exposed to high inhibitory levels of the product. In contrast, if the catalyst particles were suspended in a well mixed stirred vessel, all the catalyst will experience the same inhibitory product concentration as in the fluid stream that leaves the reactor.

6. Photobioreactors

Photobioreactors are used for photosynthetic culture of cyanobacteria, microalgae, and to a much lesser extent, cells of macroalgae (seaweeds) and plants. Photosynthesis requires light and light stimulates some cultures in ways not seen in purely heterotrophic growth. Because of the need to provide light, photobioreactors must have a high surface-to-volume ratio and this greatly affects the design of bioreactor. The demand for light is reduced in photomixotrophic culture where an organic compound is the major source of carbon for the cells and only a limited amount of photosynthesis (i.e., the fixation of carbon dioxide in presence of light) takes place.

Only a few phototrophic microorganisms mainly cyanobacteria and microalgae are cultured on large scale. This kind of mass culture is carried out in photobioreactors open to atmosphere, e.g., in ponds, lagoons, and "raceway" channels (Fig. 9). The latter are widely used and

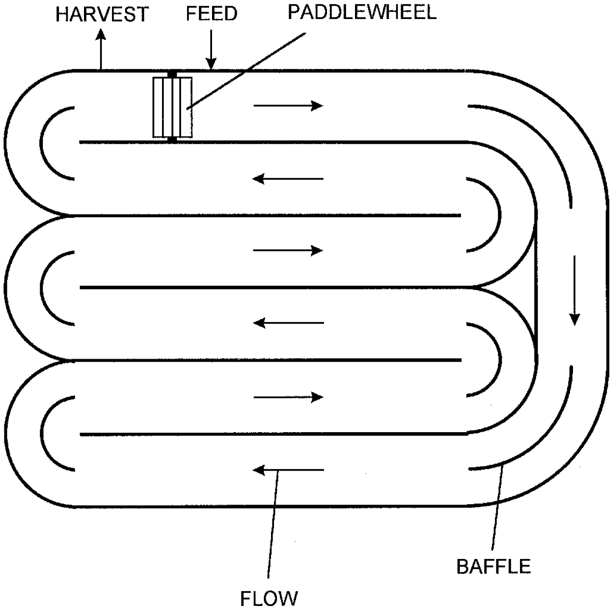


FIGURE 9 A closed-loop raceway channel for outdoor culture of photosynthetic microorganisms.

they consist of a closed loop recirculation channel that is about 0.4 m or less in depth. The culture is mixed and circulated by a paddle wheel (Fig. 9) or pumps. The channel is built in concrete and may be lined with plastic. The culture feeding and harvest are continuous, except during the night; however, the channel keeps circulating even during the night. Use of open photobioreactors is limited to only a few microbial species—the astaxanthin producer *Dunaliella*, *Chlorella*, and *Spirulina*. The last two are used mostly as healthfood. These few species can be grown in selective environments (e.g., highly alkaline and saline) that suppress contamination by other microorganisms.

A greater variety of algae and tissue cells may be cultured in fully closed photobioreactors; however, like the open raceways and ponds, a closed photobioreactor must have a high surface-to-volume ratio to effectively capture light and this requirement greatly increases the installation and operational expenses of these systems. Closed photobioreactors are used mainly to produce biomass for aquaculture feeds. A closed photobioreactor consists of a light capture unit or photoreceiver, a pumping device to circulate the culture through the photoreceiver, and a gas-exchange column to remove the photosynthetically generated oxygen and provide the carbon source (carbon dioxide). Most useful photoreceivers or photocollectors are made of ridged or flexible plastic tubing. Glass tubing is used in some cases.

The tubes may be arranged into a ladder configuration (Fig. 10a) that may be installed flat on the ground or it may be positioned at a 90° angle, as a fence. A continuous run tube may be formed into a serpentine configuration (Fig. 10b) or it may be made into a "biocoil." The latter is obtained by helically winding a flexible polymer tubing on a cylindrical-shaped frame made of aluminum or other metal bars (Fig. 10c). The culture tubes are 3 cm in diameter. The continuous run length of a single tube depends on the oxygen generation rate and the culture velocity. The length of a single tube is usually 50 m, or less. Several sections of continuous run tubing are installed on a given frame and common headers are used to distribute and collect the broth. The height of the coil may be up to 8 m and the diameter may be 2 m or more. (A larger diameter improves illumination in the region enclosed by the coil. The optimum diameter depends on the height of the coil.) The culture is circulated by a pump or an airlift device. The airlift column or a separate tower is used for gas-exchange. Temperature is controlled by evaporative cooling of water sprayed on the solar receiver. Also, a tubular heat exchanger may be used instead of evaporative cooling.

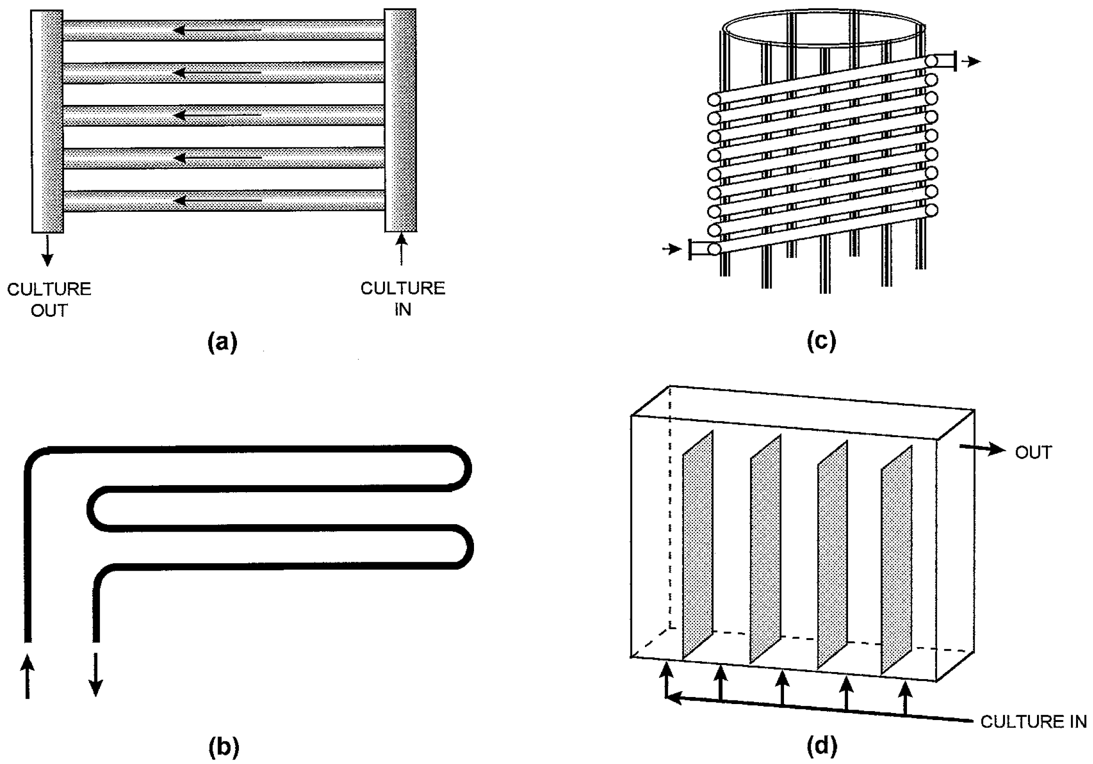


FIGURE 10 Light capture systems for photobioreactors: (a) tubular ladder, (b) continuous run serpentine, (c) biocoil, and (d) thin channels.

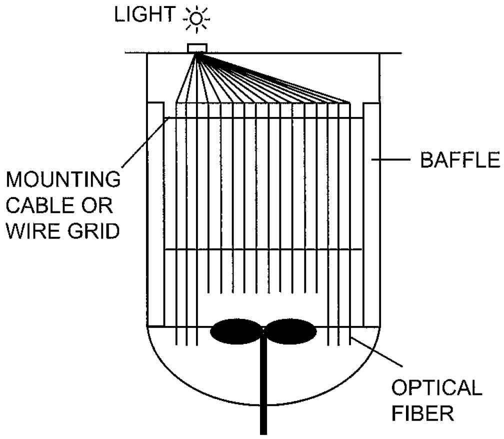


FIGURE 11 Illumination using optical fibers or waveguides.

Instead of using tubes, the photoreceiver is sometimes made of transparent plastic sheets, as in Fig. 10d. In other cases, a conventional vessel with a low surface-to-volume ratio may be illuminated by using optical fibers to convey light inside from an external source (Fig. 11), but this arrangement is not particularly effective. Irrespective of the design of the photoreceiver and the source of illumination (natural or artificial), light is generally the limiting nutrient in phototrophic culture. Except in optically dilute cultures, exponential growth does not persist for long in photosynthetic microbial culture. Because of light absorption and self-shading by cells in dense culture, light soon become limiting and growth kinetics change from exponential to linear. The depth related decline in light intensity is governed by the Beer–Lambert relationship, as follows:

$$\frac{I}{I_o} = \exp(-K_a X L), \qquad (1)$$

where I_o is the incident light intensity, I is the intensity at depth L, X is the biomass concentration, K_a is the light absorption or extinction coefficient that depends on the pigment content of the cells, and L is the culture depth. Obviously, culture depth (i.e., tube diameter, channel depth) must remain quite shallow, or the local light intensity will become too low to support growth.

7. Other Bioreactor Configurations

The basic bioreactor configurations discussed above for heterotrophic growth (i.e., stirred tanks, bubble columns and airlift bioreactors, packed and fluidized beds) are generally satisfactory for a great majority of bioprocessing needs. In addition, some basic configurations have been especially adapted to better suite specific applications. For example, stirred vessels for animal and plant cell cultures employ different designs of impeller compared to ones used in typical microbial culture. Similarly, some animal cell culture vessels are installed with bundles of microporous polymer tubing for bubble-free oxygen supply to animal cells that are particularly susceptible to damage by bursting bubbles. Oxygen or a gas mixture containing oxygen flows through the tubing and oxygen diffuses into the culture broth through the liquid film held within pores in the tube wall. In other instances, a water-immiscible oxygen carrying liquid, or an oxygen vector (e.g., perfluorocarbons and silicone oils), is used to supply oxygen, as shown in Fig. 12. The carrier fluid is oxygenated in a sparged column by bubbling with air. The bubble-free carrier then circulates through the culture vessel where oxygen is transferred to the broth. The oxygen depleted carrier loaded with carbon dioxide returns to the aeration column (Fig. 12).

Other bioreactor designs include the rotating drum fermenter (Fig. 13) with internal baffles. This device is used to culture some suspended plant cells. The drum is filled to less than 40% of its volume and rotated on rollers for mixing. Another bioreactor configuration that is suitable for hairy root cultures of plants is the mist, spray, or fog bioreactor. The static root mass is contained in a chamber that is mostly empty. In this design, the nutrients are supplied as a mist of fine droplets suspended in circulating air currents that penetrate the spaces between the roots. The spray of nutrient solution is produced by using a compressed gas atomizer nozzle or a spinning disc spray device.

B. Solid-State Culture

Solid-state culture differs markedly from submerged culture. The substrates of solid-state fermentations are particulate solids that contain little or no free water. Steamed rice is a typical substrate. Beds of solids are difficult to agitate and solid-state fermentations do not employ intensive mixing. Small particles with large surface-to-volume ratios are preferred substrates because they present a larger surface for microbial action. However, particles that are too small, and shapes that pack together tightly (e.g., flat flakes, cubes), are undesired because close packing reduces interparticle voids that are essential for aeration. Similarly, too many fines in a batch of larger particles will fill up the voids. For fermentation, the substrate is loosely packed into shallow layers or heaps. Deep beds of substrate require forced aeration with moistened air. Aeration rates may vary widely; a typical range being (0.05–0.2) $\times 10^{-3}$ m^3 kg^{-1} min^{-1}. Occasional turning and mixing improve oxygen transfer, and reduce compaction and mycelial binding of substrate particles.

Unlike many submerged fermentations, solid-sate processes commonly use mixed cultures. Hygienic processing practices are followed in large industrial operations, but

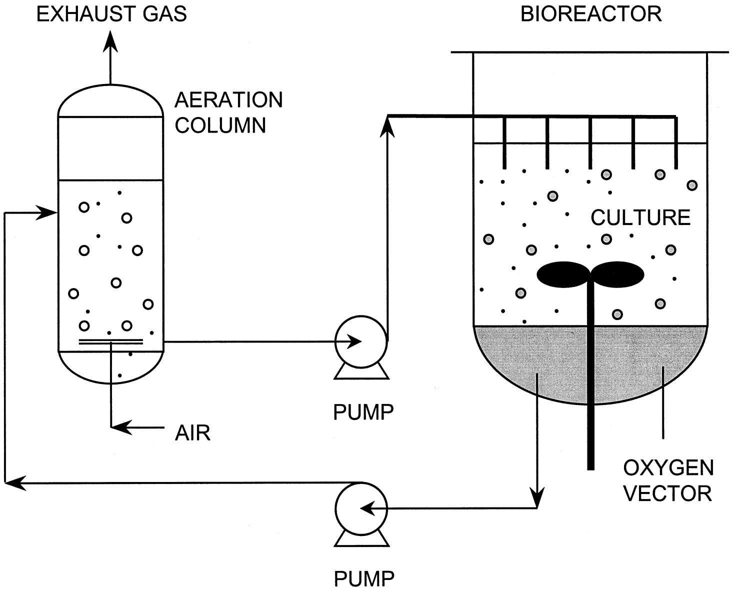


FIGURE 12 Bioreactor system for bubble-free aeration through a water-immiscible oxygen vector.

sterility standards that are common in submerged culture production of pharmaceuticals are not attained. Solid-state fermentation devices vary in technical sophistication from the very primitive banana leaf wrappings, bamboo baskets, and substrate heaps to the highly automated *koji* machines used mainly in Japan. Koji fermentations are widely practiced for making foods such as soy sauce. Koji, or molded grain, is a source of fungal enzymes that digest proteins, carbohydrates, and lipids into nutrients used by other microorganisms in subsequent fermentation. Koji comes in many varieties depending on the mold, the substrate, the method of preparation, and the stage of harvest.

Koji for soy sauce is made from soybeans and wheat. Soybeans, defatted soybean flakes, or grits are moistened and cooked in continuous pressure cookers. Cooked beans are mixed with roasted, cracked wheat. The mixed substrate is inoculated with a pure culture of *Aspergillus oryzae* (or *A. sojae*). The fungal spore density at inoculation is about 2.5 $\times 10^8$/kg of wet solids. After a 3-day fermentation the substrate mass becomes green–yellow because of sporulation. Koji is now harvested for use in a second submerged fermentation step. Koji production is highly automated and continuous. Processes producing up to 4,150 kg h^{-1} koji have been described. Similar large-scale operations are used also to produce koji for miso and sake.

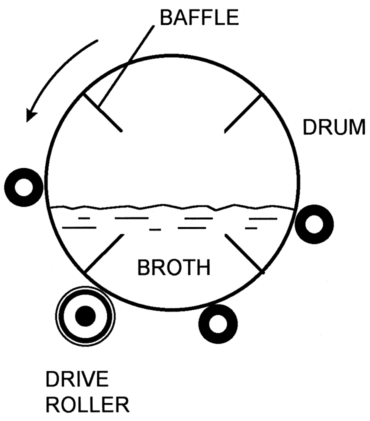


FIGURE 13 Rotating drum bioreactor for submerged culture.

Some common types of commercial solid-state fermenters are shown in Fig. 14. The *static bed fermenter* employs a single, static bed of substrate (e.g., steamed rice) located in an insulated chamber (Fig. 14a). The depth of the bed is usually less than 0.5 m. Oxygen is supplied by forced aeration through the substrate. The *tunnel fermenter* is an adaptation of the static bed device (Fig. 14b). Typically, the bed of solids is quite long, but again no deeper than 0.5 m. Tunnel fermenters may be highly automated with mechanisms for mixing, inoculation, continuous feeding, and harvest of substrate. The *agitated tank fermenter* uses one or more helical screw agitators mounted in a cylindrical or rectangular tank to agitate the fermenting substrate (Fig. 14c). The screws may move on horizontal rails. The *rotary drum fermenter* consists of a cylindrical drum that is supported on rollers (Fig. 14d). The rotation (1–5 rpm) of the drum causes a tumbling movement of the solids inside. Tumbling is aided by straight or curved baffles attached to the inside walls (Fig. 14d). Sometimes the drum may be inclined, causing the substrate to move from the higher inlet end to

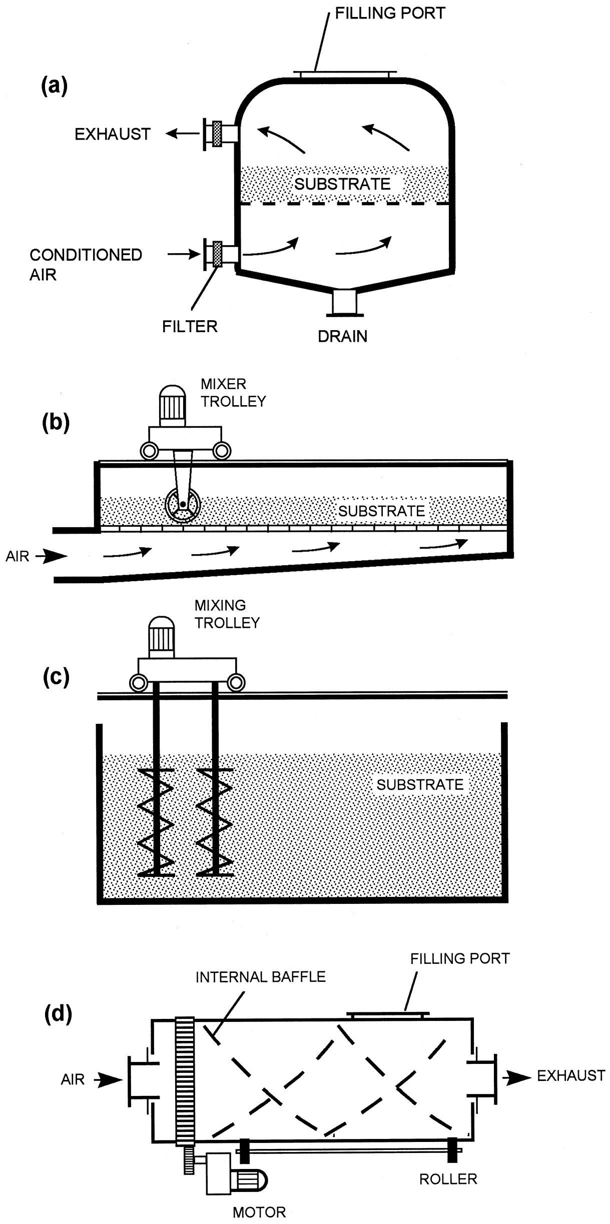


FIGURE 14 Bioreactors for solid-state fermentations: (a) static bed fermenter, (b) tunnel fermenter, (c) agitated tank fermenter, (d) rotary drum fermenter.

the lower outlet during rotation. Aeration is through the coaxial inlet and exhaust nozzles.

C. Bioreactors for Immobilized Enzymes and Cells

Immobilized enzyme and cell particles may be used in packed bed bioreactors, or the particles may be suspended in stirred tanks, bubble columns, airlift bioreactors, and fluidized beds, as discussed in an earlier section of this article. Monolayers of animal cells anchored on small spherical microcarrier particles suspended in stirred bioreactors are widely used in producing viral vaccines (e.g., rabies and polio) and some therapeutic proteins. Similarly, suspended inert particles carrying microbial biofilms are employed in some wastewater treatment processes and other fermentations. Microbial biofilms growing on rotating discs (Fig. 15) are also used to treat wastewater. The discs, mounted on a shaft, slowly rotate through a pool of the water being treated. The dissolved pollutants in the water are taken up by the cells and degraded. The oxygen needed for the degradation diffuses into the biofilm as the disc cycles through the atmosphere.

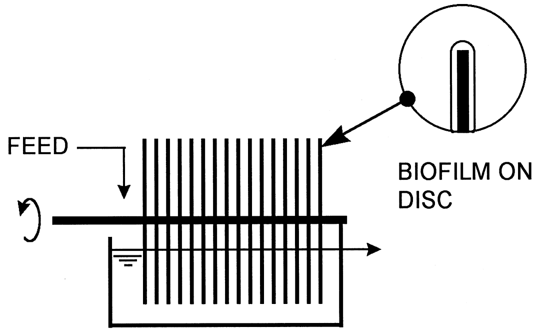


FIGURE 15 Rotating disc biofilm reactor. Microorganisms grow attached to the surfaces of the discs that rotate slowly to provide the cells with dissolved nutrients and oxygen.

Other bioreactor configurations have been developed specifically for immobilized enzymes and cells. Enzymes immobilized within polymeric membranes are used in hollow fiber (Fig. 16) and spiral membrane bioreactors (Fig. 17). In the hollow fiber device, many fibers are held in a shell-and-tube configuration (Fig. 16) and the reactant solution (or feed) flows inside the hollow fibers. The permeate that has passed through the porous walls of the fibers is collected on the shell side and contains the product of the enzymatic reaction. Also, instead of being immobilized in the fiber wall, enzymes bound to a soluble inert polymer may be held in solution that flows inside the hollow fiber. The soluble product of the reaction then passes through the fiber wall and is collected on the shell side; the enzyme molecule, sometimes linked to a soluble polymer, is too large to pass through the fiber wall.

Hollow fiber modules are sometimes used to culture animal cells that are confined to the shell side of the module; the separately oxygenated nutrient solution flows inside the fibers and perfuses the cells on the shell side. In the spiral membrane reactor (Fig. 17), the membrane that contains the immobilized enzyme is rolled into a spiral and confined within a shell. The feed or reactant solution flows in at one end and the product is removed from the opposite end of the cylindrical shell.

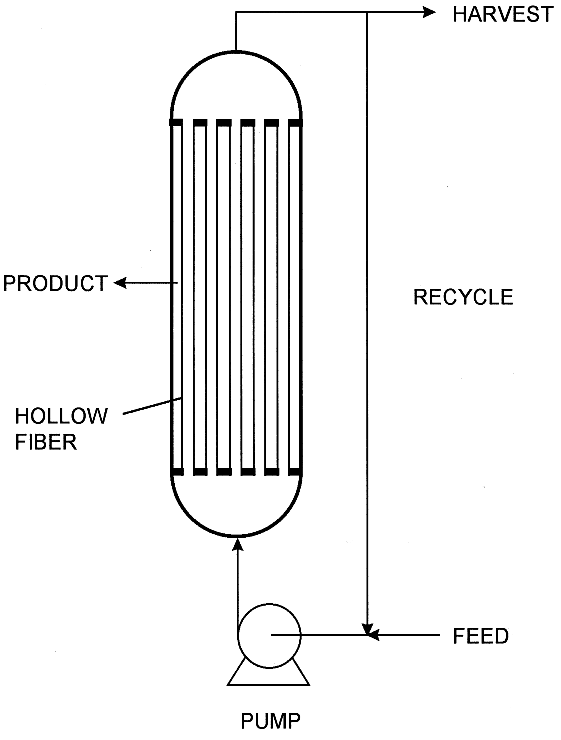


FIGURE 16 Hollow fiber membrane bioreactor.

III. CONSIDERATIONS FOR BIOREACTOR DESIGN

A. General Features

All bioreactors for monoseptic submerged culture have certain common features, as shown in Fig. 18. The reactor vessel is provided with side ports for pH, temperature, and dissolved oxygen sensors. Retractable sensors that can be replaced during operation are used commonly. Connections for acid and alkali (for pH control), antifoam agents,

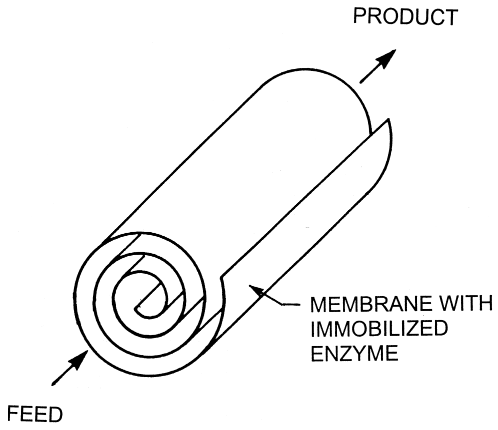


FIGURE 17 Spiral membrane bioreactor.

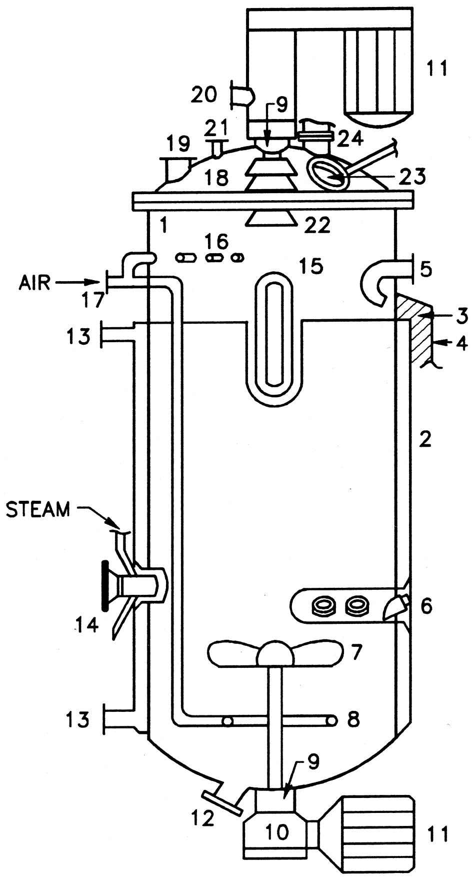


FIGURE 18 A typical submerged culture fermenter: (1) reactor vessel; (2) jacket; (3) insulation; (4) protective shroud; (5) inoculum connection; (6) ports for pH, temperature, and dissolved oxygen sensors; (7) agitator; (8) gas sparger; (9) mechanical seals; (10) reducing gearbox; (11) motor; (12) harvest nozzle; (13) jacket connections; (14) sample valve with steam connection; (15) sight glass; (16) connections for acid, alkali, and antifoam agents; (17) air inlet; (18) removable top; (19) medium feed nozzle; (20) air exhaust nozzle (connects to condenser, not shown); (21) instrumentation ports for foam sensor, pressure gauge, and other devices; (22) centrifugal foam breaker; (23) sight glass with light (not shown) and steam connection; (24) rupture disc nozzle.

and inoculum are located above the broth level in the reactor vessel. The liquid level can be easily seen through a vertical sight glass located on the vessel's side (Fig. 18). A second sight glass is located on the top of the vessel and an externally mounted light can be used to illuminate the inside of the bioreactor. The sight glass on top can be internally cleaned by a jet of steam condensate. The vessel may be placed on a load cell to obtain a better indication of the amount of material it contains.

When mechanical agitation is used, either a top or bottom entering agitator may be employed. The bottom entry design is more common and it permits the use of a shorter agitator shaft, often eliminating the need for support

bearings inside the vessel. The shaft of the agitator is provided with steam sterilizable single or double mechanical seals. Double seals are preferred, but they require lubrication with cooled clean steam condensate, or other sterile fluid. Alternatively, when torque limitations allow, magnetically coupled agitators may be used thereby eliminating the mechanical seals.

An air (or other gas mixture) sparger supplies oxygen (and sometimes carbon dioxide or ammonia for pH control) to the culture. Aeration of fermentation broth generates foam. Typically, 20 to 30% of the fermenter volume must be left empty to accommodate the foam and allow for gas disengagement. Foaming in bioreactors is controlled by a combination of chemical and mechanical methods. Chemical antifoaming agents are commonly mixed with the broth at initiation of fermentation. Further additions of antifoam agent are made from time to time, as needed. Typical antifoams are silicone oils, vegetable oils, and substances based on low molecular weight poly(propylene glycol) or poly(ethylene glycol). Emulsified antifoams are more effective because they disperse better in the fermenter. Excessive use of antifoams may interfere with some downstream separations such as membrane filtrations. Hydrophobic silicone antifoams are particularly troublesome, as they foul membrane filters and chromatography media. The use of antifoam chemicals is minimized by combining it with mechanical breakage of foam. A mechanical "foam breaker" may be installed in the headspace of the fermenter, as shown in Fig. 18. The device in Fig. 18 separates the foam—a dispersion of gas in liquid—into its components by centrifugal action, as explained in Fig. 19. The operation of the foam breaker and the addition of antifoam chemicals are controlled by signals from a foam sensor that extends into the bioreactor from the top. The shaft of the high-speed mechanical foam breaker must also be sealed using double mechanical seals as explained for the agitator.

In most instances, the bioreactor is designed for a maximum allowable working pressure of 3.78–4.10 bar (absolute pressure) at a design temperature of 150–180°C. The vessel is designed to withstand full vacuum. In North America the design conforms to the American Society of Mechanical Engineers (ASME), Section VIII, Division 1, *Boiler and Pressure Vessel Code*. Other codes may be acceptable in other locations. The reactor can be sterilized in place using saturated clean steam at a minimum pressure of 2.1 bar (absolute pressure). Overpressure protection is provided by a rupture disc located on top of the bioreactor. The rupture disc is piped to a contained drain. Usually a graphite burst disc is used because it does not crack or develop pinholes without failing completely. Other items located on the head plate of the vessel are nozzles for media or feed addition and for sensors (e.g., the foam electrode),

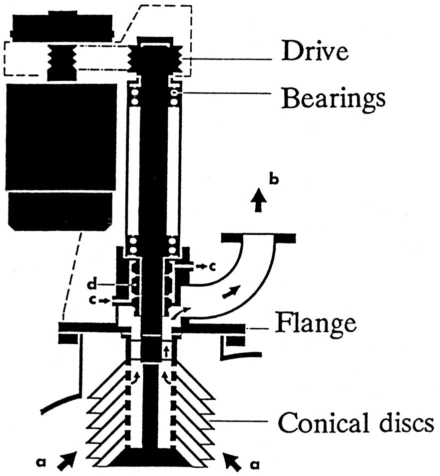


FIGURE 19 A mechanical foam breaker. The motor, drive, and shaft assembly are used to rotate the stack of conical discs at a high speed. The foam enters the spaces between the rotating discs at **a** and is separated into gas and liquid by the centrifugal force. The liquid spins into the bioreactor and liquid-free gas exhausts through the nozzle **b**. The mechanical seal **d** prevents leakage into and out of the sterile bioreactor. The seal is lubricated by sterile cooling water **c**.

and instruments (e.g., the pressure gauge). The vessel is designed to drain completely and a harvest nozzle is located at the lowest point on the reactor vessel (Fig. 18). The reactor is either provided with a manhole, or the top is removable. Flat head plates are commonly used in smaller vessels, but a domed construction of the head is less expensive for larger bioreactors (Fig. 18).

The bioreactor vessel should have few internals; the design should take into account the clean-in-place and sterilization-in-place needs. There should be a minimum number of ports, nozzles, connections, and other attachments consistent with the current and anticipated future needs of the process. The bioreactor should be free of crevices and stagnant areas where pockets of liquids and solids may accumulate. Attention to design of such apparently minor items as the gasket grooves is important. Easy to clean channels with rounded edges are preferred. As far as possible, welded joints should be used in preference to sanitary couplings. Steam connections should allow for complete displacement of all air pockets in the vessel and associated pipework, for sterilization. Evan the exterior of a bioprocess plant should be cleanly designed with smooth contours, minimum bare threads, and so forth.

The reactor vessel is invariably jacketed. In the absence of especial requirements, the jacket is designed to the same specifications as the vessel. The jacket is covered with chloride-free fiberglass insulation which is fully enclosed in a protective shroud as shown in Fig. 18. The jacket is provided with overpressure protection through a relief valve located on the jacket or its associated piping.

For a great majority of applications, austenitic stainless steels are the preferred material of construction for bioreactors. The bioreactor vessel is usually made in Type 316L stainless steel, while the less expensive Type 304 (or 304L) is used for the jacket, the insulation shroud, and other non-product contacting surfaces. The L grades of stainless steel contain less than 0.03% carbon, which reduces chromium carbide formation during welding and lowers the potential for later intergranular corrosion at the welds. The welds on internal parts should be ground flush with the internal surface and polished. Welds are difficult to notice in high-quality construction. In addition to the materials of construction, the surface finish also requires attention. The finish on surfaces which come in contact with the product material and, to some extent, the finish on external surfaces affects the ability to clean, sanitize, and sterilize the bioreactor and the general processing area. The surface finish has implications on stability and reactivity of the surface, and it may have process implications relating to microbial or animal cell adhesion to surfaces.

The mill-finished surface of stainless-steel sheet is unsatisfactory for use in bioreactors. Minimally, the surface should receive a mechanical polish. Mechanical polish is achieved by abrasive action of a sandpaper type material on metal. The surface finish may be specified by grit number, for example, 240-grit polish, which refers to the quantity of particles per square inch of the abrasive pad. The higher the grit number, the smoother the finish. More quantitative measures of surface finish rely on direct measurement of roughness in terms of "arithmetic mean roughness," Ra, or "root mean square roughness." Microscopic examination of even a highly smooth mechanically polished surface reveals a typical pattern of grooves and ridges that provide sites for microbial attachment. For example, a 320-grit polished surface will have an Ra of the order of 0.23–0.30 μm. Hence, for internal surfaces of bioreactors, electropolished surface is preferable to mechanical polish alone.

Electropolishing is an electrolytic process which preferentially removes the sharp microscopic surface projections arising from mechanical polishing; the result is a much smoother finish. Electropolishing significantly reduces the metal surface area and, hence, the product-metal contact area. The treatment imparts corrosion resistance to stainless steel by removing microscopic regions of high local stress; it creates a passivated steel surface, rich in protective chromium oxide. To attain a suitable electropolished finish, the surface should be previously mechanically polished; however, there is little advantage to starting with a much better than 220-grit (Ra $\approx$ 0.4–0.5 μm) polished surface. If mechanical polish alone must be used, it should be at least 240 grit, and the direction of polish should be controlled to produce a vertical grain for good drainage. The surface should receive a nitric acid wash treatment as a minimum. The orientation of the grain does not seem to be of consequence if the surface is to be electropolished.

B. Mixing, Heat, and Mass Transfer

1. Mixing and Shear Effects

A minimal intensity of mixing is required in a bioreactor to suspended the biocatalyst and substrate particles, prevent development of pH and temperature gradients in the bulk fluid, and improve heat and mass transfer. Mixing also enhances transfer of nutrients and substrates from the fluid to the biocatalyst particles and helps remove and dilute inhibitory metabolites that may be produced. Mixing is generally provided by mechanical agitation or by bubbling compressed gas into the fluid. Excessively intense mixing is harmful; too much turbulence damages certain cells, disintegrates immobilized biocatalyst pellets, and may dislodge biofilms from carriers. Freely suspended microorganisms are generally tolerant of hydrodynamic forces (or "shear" forces) encountered in bioreactors under typical conditions of operation; however, animal cells, suspended plant cells, certain microalgae, and protozoa are especially prone to shear damage. Forces associated with rupture of sparged gas bubbles are known to destroy animal cells. Damage is minimized by using lower aeration velocities, larger bubbles (diameter $\geq$ 0.01 m), and supplementation of the culture medium with protective additives such as the surfactant Pluronic F68. Aeration associated power input in bioreactors for animal cell culture is typically kept at less than 50 W m^{-3}. The power input may be calculated as follows:

$$\frac{P_G}{V_L} = \rho_L g U_G, \tag{2}$$

where P_G is the power input, V_L is the volume in the bioreactor, ρ_L is the density of the broth, g is the gravitational acceleration, and U_G is the superficial gas velocity. The velocity is calculated as the volume flow rate of the gas divided by the cross sectional area of the bioreactor.

2. Oxygen Supply and Carbon Dioxide Removal

Animal and plant cells need oxygen to survive. Many microorganisms require oxygen (i.e., they are *obligate aerobes*) but oxygen may be toxic to others (*anaerobes*). Some microbes may switch between aerobic and anaerobic growth and are said to be *facultative*. Sufficiency of oxygen supply is necessary to prevent growth limitation in aerobic cultures. Oxygen is provided usually by sparging the broth with air or some other oxygen-containing mixture of gases. Other specialized methods

of oxygen supply are used in a few bioprocesses (e.g., Fig. 12).

Oxygen is sparingly soluble in aqueous broths and even a short interruption in aeration rate in some microbial fermentations may produce anaerobic conditions that may potentially damage the cells. The precise oxygen requirements of a fermentation process depend on the microorganism, the degree of oxidation of the substrate being used for growth, and the rate of oxidation. Oxygen becomes hard to supply when the demand exceeds 4–5 kg O_2 m^{-3} hr^{-1}. Many microbial fermentation broths are highly viscous and difficult to mix. This further complicates the transfer of oxygen from the gas phase to the broth. Once the concentration of dissolved oxygen falls below a critical value, the microbial growth becomes limited by oxygen. The critical dissolved oxygen concentration depends on the conditions of culture and the microbial species. Under typical culture conditions, fungi such as *Penicillium chrysogenum* and *Aspergillus oryzae* have a critical dissolved oxygen value of about 3.2×10^{-4} kg m^{-3}. For baker's yeast and *Escherichia coli*, the critical dissolved oxygen values are 6.4×10^{-5} and 12.8×10^{-5} kg m^{-3}, respectively.

Animal and plant cell cultures have lower oxygen demands than microbial cells. Oxygen consumption rates for animal cells are in the range of 0.05–0.5 mmol/10^9 cells per hr. In batch suspension culture of animal cells, the maximum cell concentration typically does not exceed 2×10^6 cells ml^{-1}. Higher concentrations, $>10^7$ cells ml^{-1}, are attained in perfusion culture without cell retention. Perfused culture with cell retention permits cell densities of around 10^9 cells ml^{-1}. *In vitro* cultured animal cells are generally tolerant of high concentrations of dissolved oxygen, e.g., up to 100% of air saturation; however, the optimal concentration is about 50% of air saturation but may vary with cell type. Oxygen concentrations of $>100\%$ of air saturation have been associated with oxidative damage to cells whereas concentrations $\leq 0.5\%$ of air saturation inhibit the TCA cycle and lead to an enhanced production of lactate. Concentrations $<10\%$ of air saturation will limit growth of some cells, but for others oxygen limitation is encountered around 0.5% of air saturation. Plant cells consume oxygen at a rate of $(3–15) \times 10^{-5}$ mol kg^{-1} DW s^{-1} and the maximum cell density in suspension culture tends to be 20–30 gDW L^{-1}.

As with dissolved oxygen, concentration of dissolved carbon dioxide influences microbial and cell growth. Too much carbon dioxide is detrimental to most aerobic fermentations but cultures of photosynthesizing microbes may require carbon dioxide in the aeration gas. Also, carbon dioxide is added at roughly 5% (by vol) to gas mixture used in sparging animal cell cultures when the culture broth is buffered with carbonate-bicarbonate system. Normal mammalian cells need carbon dioxide as a substrate for carboxylation of pyruvate to oxaloacetic acid, but established cell line may not require carbon dioxide.

3. Heat Removal and Temperature Control

All fermentations generate heat. In submerged cultures, 3–15 kW m^{-3} of the heat output typically comes from microbial activity. In addition, mechanical agitation of the broth produces up to 15 kW m^{-3}. Consequently, a fermenter must be cooled to prevent temperature rise and damage to culture. Temperature is controlled by circulating cooling water in a jacket that surrounds the bioreactor vessel (Fig. 20a). In addition to the jacket, an internal cooling coil (Fig. 20b) becomes necessary in large bioreactors. In some cases, cooling is achieved by recirculating the broth through an external heat exchanger (Fig. 20c). In small vessels, a "ringlet" coil (Fig. 20d) that enters the vessel through a large port on top, may provide sufficient cooling. Heat removal tends to be difficult because, typically, the temperature of the cooling water is only a few degrees lower than that of the fermentation broth. Industrial fermentations are commonly limited by the heat transfer capability. The ability to remove heat depends on the surface area available for heat exchange, the temperature difference between the broth and the cooling water, the properties of the broth and the coolant, and the turbulence in those fluids. The geometry of the fermenter determines the heat exchange area that can be provided.

Because metabolic heat generation depends on the oxygen consumption rate, heat removal in large vessels

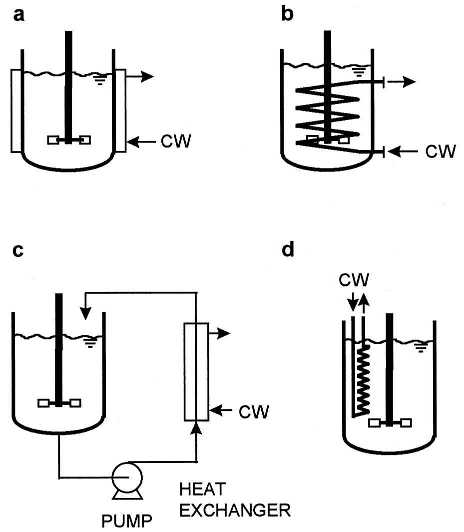


FIGURE 20 Heat-exchange methods for bioreactors: (a) jacket, (b) full internal coil, (c) recirculation of broth through an external heat exchanger, (d) ringlet coil. Cooling water (CW) enters and exits the heat exchange devices, as shown.

becomes difficult as oxygen consumption rate approaches 5 kg m^{-3} h^{-1}. About 0.54 kJ of heat is generated for each millimole of oxygen consumed. Oxygen consumption rate in microbial broths typically ranges from 2×10^{-4} to 1×10^{-3} kg m^{-3} s^{-1}. Typically, microorganisms growing on hexose sugars such as glucose produce about 10.9 MJ of heat per kilogram of biomass formed. For cells growing on highly reduced substrates such as hydrocarbons, heat generation is greater, about 28.5 MJ/kg of biomass produced. These figures are for fermentations in which biomass is the only product. Most microbial and plant cells are cultured at a temperature of between 20 and 30°C. Mammalian cells are usually grown at 37°C. Insect cells prefer a lower temperature, e.g., 26°C. Temperatures of greater than 40°C are optimal for certain *thermophilic* microorganisms.

In addition to removing the metabolic heat, a fermenter must provide for heat transfer during sterilization and subsequent cooling. Liquid (or slurried) fermentation medium for a batch fermentation may be sterilized using batch or continuous processes. With batch processes, the medium or some of its components and the fermenter are commonly sterilized together in a single step by heating the dissolved or slurried medium inside the fermenter. For *in situ* sterilization, steam may be injected directly into the medium or heating may be through the fermenter wall.

High temperature (typically 121°C) heating during sterilization often leads to undesirable reactions between components of the medium. Such reactions reduce yield by destroying nutrients or by generating growth inhibitory compounds. This thermal damage is prevented or reduced if only certain components of the medium are sterilized in the fermenter and other separately sterilized components are added later. Sugars and nitrogen-containing components are often sterilized separately. Dissolved nutrients that are especially susceptible to thermal degradation may be sterilized by passing through hydrophilic polymer filters that are rated to retain particles down to 0.45 μm. Even finer filters (e.g., 0.2 μm rated particle retention) are available.

Heating and cooling of a large fermentation batch takes time that ties up a fermenter unproductively. In addition, the longer a medium remains at high temperature, the greater is the thermal degradation or nutrient loss. Therefore, continuous sterilization of the culture medium into a presterilized fermenter is preferable even for batch fermentations. Continuous sterilization is rapid and it limits nutrient loss; however, the initial capital expense is greater because a separate sterilizer is necessary.

A heat exchange system must be able to handle the high heat loads encountered during sterilization by steam and subsequent cooling. Irrespective of the specific heat exchange system used (Fig. 20), the operational details can be quite complex as shown in Fig. 21 for a jacketed bioreactor. Valves 1, 10, and 5 in Fig. 21 are control valves that modulate the flow of steam and cooling water; valves 2 and 9 are vacuum breakers that allow the pipework to drain under gravity; valve 3 is a pressure relief valve to protect the jacket and the pipework against pressurization above the safe acceptable limit; and all other valves are pneumatically operated devices that are either fully open or fully closed.

For sterilization the jacket is heated by steam. Valves 4–6, 8, and 10 are closed and the pump 7 is off. Valve 11 is opened to drain the jacket. After a short period, valve 11 is closed and valve 12 is opened. Valve 1 is opened to let steam into the jacket. The condensate drains through the steam trap 13. The temperature and pressure of the circuit are monitored at TIC1 and PI1. Cooling is carried out by closing valves 12 and 1; valves 8 and 4 are opened. Cooling water enters the circuit at valve 4, flows up the jacket and out to the cooling water return line (Fig. 21). All operations are generally automated for consistent and error-free control.

Animal cell culture may actually require some warming to maintain the temperature at 37°C. The circuit shown in Fig. 21 uses a closed recirculation loop to control the temperature. During culture, valves 8 and 6 are opened and the pump 7 is turned on. The water is pumped through a compact plate heat exchanger (PHE) where it is indirectly heated by controlled (control valve 10) flow of steam. The heated water now passes the temperature indicator/controller TIC1 and enters the jacket. The water recirculates via valve 8 and pump 7. Cold water is injected (control valve 5) in the circuit to maintain the temperature. Any excess water leaves the loop via the relief valve 3. In some designs, direct steam injection into the circulating loop may be used instead of the heat exchanger. In small fermenters, the exchanger may be replaced by electric heating. Notice the flexible connections between the fermenter and the temperature control pipework (Fig. 21). These connections are necessary for fermenters that rest on load cells (for weight measurement); the connections allow the fermenter to move freely.

Compared to submerged culture, biomass levels in solid state fermentations are lower at 10–30 kg m^{-3}. Nevertheless, because there is little water and the density of the substrate is relatively small, the heat generation per unit fermenting mass tends to be much greater in solid state fermentations than in submerged culture. Temperature can rise rapidly, again, because there is little water to absorb the heat. Cumulative metabolic heat generation in koji fermentations for a variety of products has been noted at 419–2387 kJ kg^{-1} solids. Higher values, up to 13,398 kJ kg^{-1}, have been observed during composting. Peak heat generation rates in koji processes range over 71–159 kJ kg^{-1} hr^{-1}, but the average rates are more moderate at

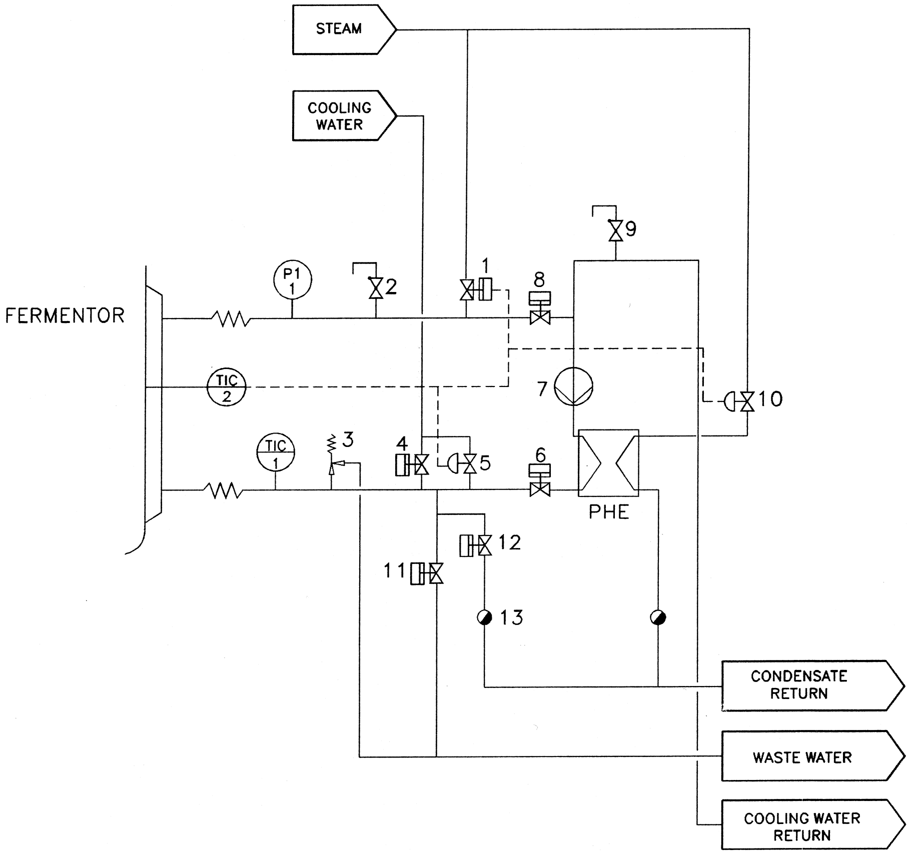


FIGURE 21 Temperature control circuit on a bioreactor.

25–67 kJ kg^{-1} hr^{-1}. Peak metabolic heat production rate during fermentation of readily oxidized substrates such as starch can be much greater than in typical koji processes.

In solid-state fermentation the substrate temperature is controlled mostly through evaporative cooling; drier air provides a better cooling effect. Intermittent spray of cool water is sometimes necessary to prevent dehydration of the substrate. Air temperature and humidity are also controlled. Occasionally, the substrate-containing metal trays may be additionally cooled by circulating a coolant, even though most relatively dry and porous substrates are poor conductors of heat. Intermittent agitation of substrate heaps further aids heat removal. Despite much effort, temperature gradients in the substrate do occur, particularly during peak growth.

C. Monoseptic Operation, Cleaning, Sterilization

Many commercial bioprocesses utilize only pure cultures. Maintenance of monoculture is vital to success of such processes; hence, a bioreactor must be sterilized prior to inoculation and contamination during operation must be prevented. A contaminating virus or microbe may destroy the culture, reduce productivity, or lead to other unwanted results. Poor design and operation of a bioreactor increase the chance of contamination and cause financial loss.

Sterilization with steam at a minimum temperature of 121°C is the norm. The sterilization temperature must be attained everywhere in the vessel and continuously held at the requisite value for at least 20 min. After sterilization and during culture the contents of the bioreactor remain fully isolated from the external environment. Air enters the vessel through presterilized hydrophobic membrane filters that prevent ingress of contaminating microbes. Similar filters are located on the air exhaust pipe. These filters are rated for removing particles down to 0.45 μm, or even 0.1 μm. The bioreactor may be sterilized together with the culture medium, or separately. In the latter case, once the bioreactor has cooled, the medium is pumped in through a sterilizing filter that removes any contaminating microbes. Alternatively, the medium is steam sterilized, cooled, and then transferred to the bioreactor through pipework that is fully isolated from the surroundings.

After the culture and before next sterilization, the bioreactor must be thoroughly cleaned usually by automated

clean-in-place methods. In-place cleaning does not require dismantling the bioreactor and saves time. Automation assures consistency of cleaning. Cleaning between batches is essential to preventing cross-contamination of products. Also, a clean bioreactor is easier to sterilize. Cleaning is achieved by physical action of high velocity flow, jet sprays, agitation, and chemical action of cleaning agents enhanced by heat. While mechanical forces are necessary to remove gross soil and to ensure adequate penetration of cleaning solutions to all areas, most of the cleaning action is provided by chemicals—surfactants, acids, alkalis, and sanitizers. A generally applicable cleaning scheme for bioreactors utilizes a water pre-rinse to remove gross soil; a hot alkali recirculation step to digest and dissolve away the remaining soil; and a water wash to remove residual alkali. A bioreactor that processes injectable drugs, should be rinsed with a hot water-for-injection (WFI) wash as the final cleaning step. Optional acid wash and sanitization steps may be added in some applications.

D. Operational Modes of Bioreactors

The way a culture is operated and fed has a profound impact on the outcome of the fermentation. Also, the reactant feeding and reactor operational strategies affect performance of enzyme bioreactors and those that use non-viable biomass as a biocatalyst. Most bioreactors for microbial growth and culture of other cells are operated as *batch* and *fed-batch* devices. A batch fermentation is initiated by inoculating a presterilized and cooled medium that is contained typically in a well-mixed bioreactor (Fig. 22a). The medium composition in the fermenter changes continuously as nutrients are consumed to produce biomass and metabolites. The broth is harvested after the designated batch time. The broth volume in a batch fermenter remains essentially constant (Fig. 22a), discounting any evaporative loss. A fed-batch culture is identical to a batch operation, except that a feed is added to the broth continuously or intermittently (Fig. 23b). The volume of the broth increases with time. Also, the feeding rate generally incenses with time, to satisfy the demand of an exponentially increasing cell population. The broth is harvested at the end of the batch period.

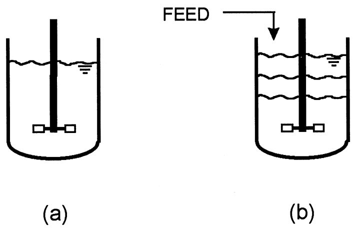


FIGURE 22 Bioreactor operational modes: (a) a well-mixed batch bioreactor, (b) a well-mixed fed-batch bioreactor.

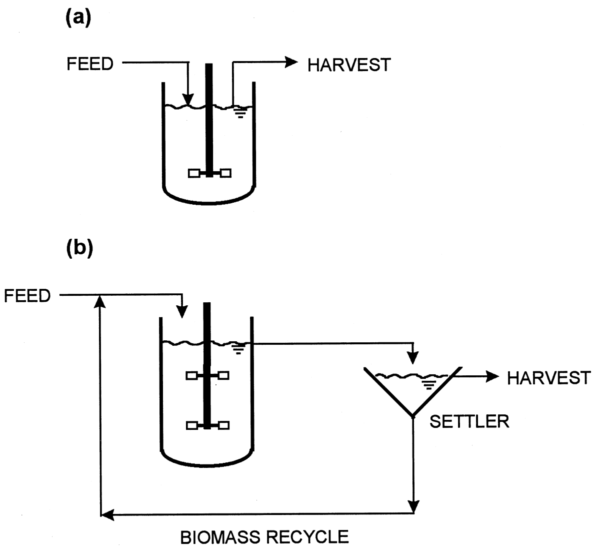


FIGURE 23 Bioreactor operational modes: (a) a continuous flow well-mixed bioreactor, (b) a continuous flow well-mixed bioreactor with biomass recycle.

In a *well-mixed steady-state continuos culture* (Fig. 23a), the broth composition does not vary with time and position in the bioreactor. Typically, a continuous fermentation starts as a batch culture and switches to continuous feeding when a sufficient concentration of biomass has been obtained. The feed is added continuously and at a constant rate. The broth volume does not change, as the rate of harvest matches the feeding rate (Fig. 23a). Sometimes, a well-mixed continuos culture may be carried out with recycle of a part of the harvested biomass to the bioreactor (Fig. 23b). This strategy increases the steady state biomass concentration in the reactor, improves conversion of the substrate, and enhances the productivity. Some continuous flow reaction and production schemes use a number of well-mixed reactors in series (Fig. 24a–c). The harvest of one reactor becomes the feed for the next. This arrangement is especially useful when different environmental conditions are needed at different stages of a process, for example, when the requirements for growing the biomass differ from the ones for synthesis of a metabolite by the cells. A multistage continuous array of well-mixed reactors may be fed with different feeds and precursor compounds at different stages, as shown in Fig. 24b. Also, such a series of reactors my employ recycle of the biomass (Fig. 24c) to the first stage, or one or more of the other stages.

Continuous flow processes sometimes use a *plug flow* bioreactor in which there is little mixing of fluid elements in the direction of flow (Fig. 25a). This type of flow is typically achieved in long tubes and channels. The composition of the broth does not change with time at a fixed

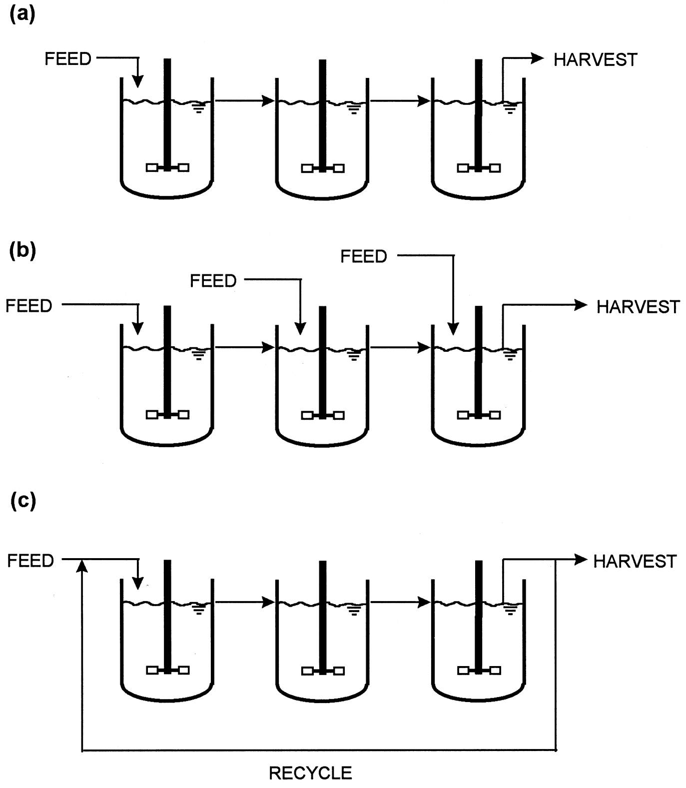


FIGURE 24 Bioreactor operational modes: (a) a series of continuous flow well-mixed bioreactors, (b) use of different feeds at different stages in a series of continuous flow well-mixed bioreactors, (c) a series of continuous flow well-mixed reactors with biomass recycle.

location in a plug flow bioreactor operating at steady state; however, the composition changes with position in the direction of flow. A plug flow bioreactor behaves the same as a series of many continuous flow well-mixed bioreactors (Fig. 24a). In terms of kinetics, a plug flow bioreactor is equivalent to a batch fermenter having the same residence time (i.e., the total time spent by the broth in the bioreactor) as in the plug flow device. A plug flow bioreactor for biomass production must be inoculated continuously, or a portion of the harvested biomass must be recycled to the inlet of the bioreactor (Fig. 25b), to prevent wash out of the culture by the sterile feed. Multipoint feeding is sometimes used in plug flow bioreactors, as shown in Fig. 25c.

Continuous flow operation of a bioreactor such that the cells are recycled or retained within the reactor, is sometimes known as perfusion culture. Many different kinds of devices are available for cell retention and recycle. These devices may be located within or outside the bioreactor vessel. External sedimentation tanks (Fig. 23b) are often used to recycle biomass in wastewater treatment processes. External and internal enhanced-rate sedimentors are also employed in animal cell culture bioreactors, as shown in Fig. 26a, b. Because of the small differences in density and diameter of the viable and nonviable animal cells, the inclined channel sedimentor (Fig. 26b) may preferentially retain viable cells in the bioreactor while allowing many of the nonviable ones to wash out. Another device that is commonly deployed for partial retention of animal cells is the "spinfilter" shown in Fig. 27. A spinfilter

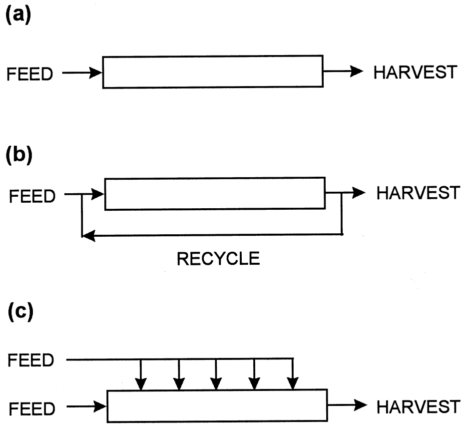


FIGURE 25 Bioreactor operational modes: (a) a continuous plug flow bioreactor, (b) a continuous plug flow bioreactor with biomass recycle, (c) multipoint feeding of a plug flow bioreactor.

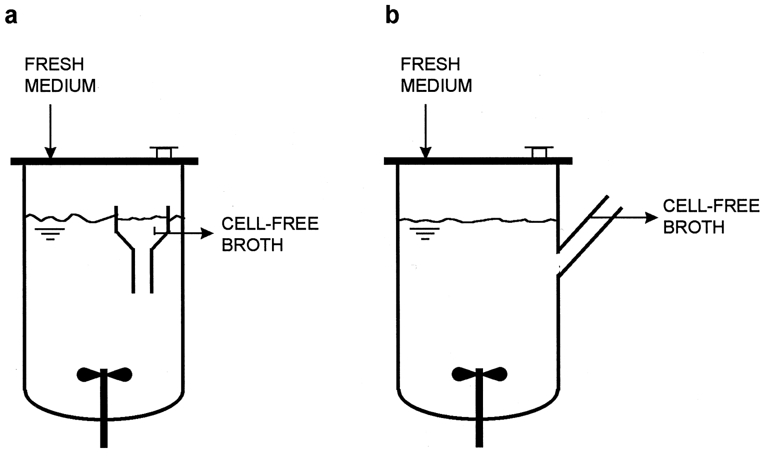


FIGURE 26 Cell retention by sedimentation in perfusion culture: (a) internal sedimentor, (b) inclined channel sedimentor.

is a cylinder formed of wire meshing. The openings of the wire screen are significantly larger (e.g., 25 μm) than the cells which are retained by a hydrodynamic mechanism requiring rapid rotation (e.g., 500 rpm) of the spinfilter. The mostly cell-free spent medium is withdrawn from the zone within the rotating screen. Depending on the specifics of design and operation, a spinfilter may preferentially retain the viable cells. Also, a noninvasive acoustic sedimentation device has been developed and this is installed on the outside of the bioreactor harvest pipe. This device uses sound waves to aggregate the cells which sediment back into the bioreactor vessel.

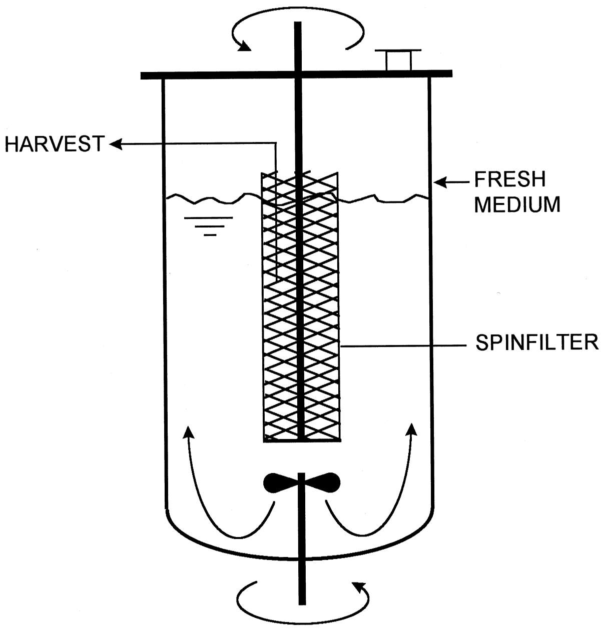


FIGURE 27 A spinfilter device of cell retention in perfusion culture of animal cells.

Spinfilters and sedimentation devices do not retain all cells within a bioreactor and some washout occurs. Total retention of biomass is feasible by using an external microfilter to continuously harvest a cell-free stream, as shown in Fig. 28. The cells are returned to the bioreactor. The bioreactor operational schemes identified here are the ones most commonly used; other variations on these schemes are encountered occasionally.

E. Medium Composition

The composition of the nutrient medium used to grow cells determines the rate of production, the type of products produced, and the yield of biomass and products. The medium also influences the cost of production, the quality of the product, and factors such as the broth rheology. The medium must provide all the elements that compose the product and the biomass. These elements (e.g., C, N, O, S, P) must be provided in a suitable form and ratios that are designed to achieve specific effects. The growing cells may require additional complex organic molecules (micronutrients) that they are unable to synthesize but that are essential to growth. In addition, the medium may contain specific precursor compounds, the substrates that are being biotransformed, and inhibitors or inducers of specific enzymes. While the medium needs many components, care is necessary to prevent contamination with others that may affect the process performance. For example, presence of iron in *Aspergillus niger* fermentation broth greatly suppresses the production of citric acid. Proper formulation of the medium is also important in cell-free enzyme reactors, where the rate of reaction and the extent of substrate (or/and product) inhibition depend on the composition of the reaction medium. Some enzyme reactions are carried out in organic solvents containing small amounts of water.

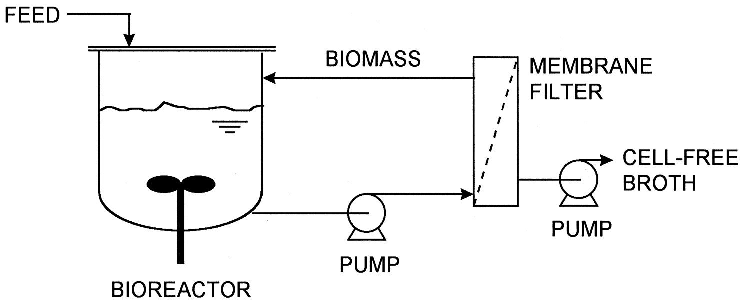


FIGURE 28 Total recycle of biomass by using a microfiltration device.

Medium design for viable culture needs to consider the osmotic pressure of the broth as an important variable.

Osmotic pressure of a solution is a measure of the concentration of dissolved molecules and ions present as individual particles. All cells are susceptible to osmotic pressure of the surrounding medium, but cells without walls (e.g., animal cells, protoplasts) are especially susceptible to osmotic stress related damage. When the suspending fluid has a higher osmotic pressure than the cells, water is drawn out of the cells and it may become so dehydrated that metabolism ceases. Incidentally, this is why high salt and high sugar media often have a preservative action, e.g. in pickling of vegetables. When the osmotic pressure is low compared to that in the cell, the cell takes up water and may burst. Osmotic shock produced by rapid dilution is sometimes used to rupture cells, especially animal cells. Osmotic pressure also influences plant cell suspension culture.

Media formulations do not generally specify the osmotic pressure; instead, the number of dissolved particles is given as osmolarity or osmolality. Osmolality is the number of moles of particles per kilogram of solution whereas osmolarity is the number of moles per liter of solution. One mole of particles is an osmole, abbreviated as Osm. Animal cell culture media have an osmolality of 280 to 320 mOsm kg^{-1} to conform to the osmolality of serum (290 mOsm kg^{-1}). Osmolality is not easily calculated especially when a medium has many components and the degree of dissociation is not known. In such cases, osmolality is estimated from measurements of freezing point depression and other colligative properties (i.e., those dependent on the concentration of dissolved particles). In cell culture media, sodium chloride is used for adjusting osmolality to the requisite value.

F. Kinetics, Productivity, and Bioreactors

Design and performance analysis of a bioreactor are inseparably linked with the kinetics of the bioreaction for which the reactor is intended. Some common bioreactions include growth of microorganisms and other cells, and reactions involving cell-free enzymes. Essential aspects of kinetics for bioreactor design are discussed here.

1. Cell Growth

Once a properly formulated and sterilized medium has been inoculated with a *seed culture* or inoculum, the cells grow and multiply. In a batch culture, the cell or biomass concentration increases with time as shown in Fig. 29, which is typical. A short *lag phase* of little or no growth is followed by a period of *exponential growth*. The lag phase is an adaptation period in which the cells become acclimatized to a new environment. The length of the relatively unproductive lag phase may be shortened by increasing the inoculum size and ensuring that the growth environment (medium, pH, temperature, etc.) in the bioreactor is the same as the one in which the inoculum was grown. Typically, the inoculum should be in the late exponential phase of growth. The volume of a microbial inoculum should between 5 and 10 percent of the volume of the medium being inoculated. Larger inocula are needed for slower growing cells such as animal and plant cells. In cultivation of animal cells, the inoculum size is selected to provide an initial cell count of (2–4) $\times 10^5$ cells/ml. After the lag period, the exponential growth persists usually until an essential nutrient (the growth limiting nutrient or substrate) runs low in the medium, or until some inhibitory product of metabolism accumulates to a growth inhibitory concentration. The cells then enter a phase of zero net growth, or the *stationary phase* (Fig. 29). This is followed by a *decline phase* or death phase in which there is a net loss of biomass as cells die and lyse. Usually an attempt is made to achieve a maximum growth rate and maintain it for as long as possible.

The biomass growth rate in a bioreactor, i.e., the change in biomass concentration with time, or dX/dt, depends on the viable biomass concentration X present at any time. In other words, growth is self catalyzing, or *autocatalytic*. In exponential growth, the growth rate is expressed as follows:

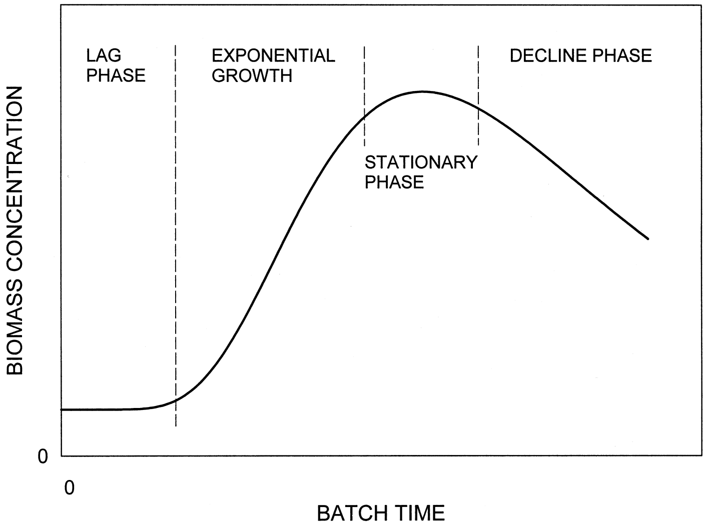


FIGURE 29 A typical biomass growth profile in batch culture.

$$\frac{dX}{dt} = \mu X, \tag{3}$$

where μ is a constant known as the *specific growth rate*. Equation (3) may be written in its integrated form, as follows:

$$\ln \frac{X}{X_o} = \mu t, \tag{4}$$

where X_o is the initial biomass concentration at any time $t=0$ and X is the concentration at time t. The time to double the biomass concentration, or the *doubling time* t_d, may be estimated from Eq. (4) by substituting t_d for t and $2X_o$ for X_o; thus,

$$t_d = \frac{\ln 2}{\mu}. \tag{5}$$

Some typical values of the doubling time are noted in Table I for the various kinds of microbial and other cells.

In steady-state continuous culture in which the cells experience an unchanging environment, the specific growth rate depends on the concentration S of a *growth limiting substrate*. This dependence is generally described by Monod kinetics, as follows:

TABLE I Typical Doubling Times

Cell type	t_d (min)
Bacteria	20–45
Yeasts	90
Molds	160
Protozoa	360
Hybridomas	630–1260
Plant cells	3600–6600

$$\mu = \frac{\mu_{max} S}{K_s + S}, \tag{6}$$

where μ_{max} is the *maximum specific growth rate* and K_s is the value of S at which the specific growth rate is half of its maximum value. K_s is known as the *saturation constant*. The specific growth rate increases with increasing substrate concentration in a hyperbolic manner, as shown by the solid line in Fig. 30. The figure also clarifies the meanings of μ_{max} and K_s.

Sometimes, the growth rate is suppressed in the presence of too much substrate and growth is said to be *substrate inhibited*. In substrate inhibited culture, the specific growth rate attains a maximum value as the substrate concentration is increased and then the growth rate declines, as shown by the dashed line in Fig. 30. Note that in Fig. 30, the μ_{max} values are different for the two growth profiles shown. The substrate inhibited growth may be described by the following equation:

$$\mu = \frac{\mu_{max} S}{K_s + S + \dfrac{S^2}{K_i}}, \tag{7}$$

where K_i is the *inhibition constant*. In other cases, the growth rate may be subject to inhibition by a product of metabolism.

2. Productivity

Productivity of a bioreactor is the quantity of product produced per unit volume in unit time. For a batch bioreactor with the growth profile shown in Fig. 29, the biomass productivity P at any time t is the slope of the straight line

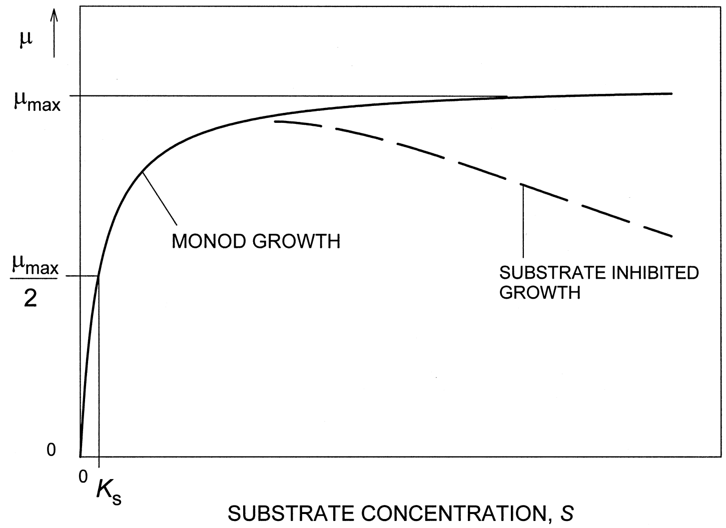


FIGURE 30 Dependence of the specific growth rate μ on the concentration S of the growth limiting substrate.

joining the biomass concentrations at $t = t$ and $t = 0$ in Fig. 31; i.e.,

$$P = \frac{\Delta X}{\Delta t}. \quad (8)$$

This definition of productivity disregards any time invested in preparing the fermenter prior to inoculation of the batch. Lower values of productivity result when the additional preparatory time is taken into account. Often, the product of a fermenter is not the biomass per se, but a compound produced by the cells. Examples of such products are antibiotics and monoclonal antibodies. The kinetics of product formation may mirror those of the cell growth, or they may be quite different. Products known as *secondary metabolites* (i.e., ones that are nonessential to the cell) are often produced after the growth has ceased.

An important operational variable for continuous flow bioreactors is the *dilution rate* D, or the volume flow rate of the feed medium divided by the constant volume of the broth in the reactor. The biomass productivity of a continuous culture is simply the dilution rate D multiplied by the biomass concentration X in the harvest stream. In a continuous flow well-mixed bioreactor, the biomass concentration varies with dilution rate, as shown in Fig. 32. Also, the productivity increases with increasing dilution rate until the dilution rate approaches close to the maximum specific growth rate (Fig. 32). Any further increase in dilution rate causes a sharp decline in productivity and

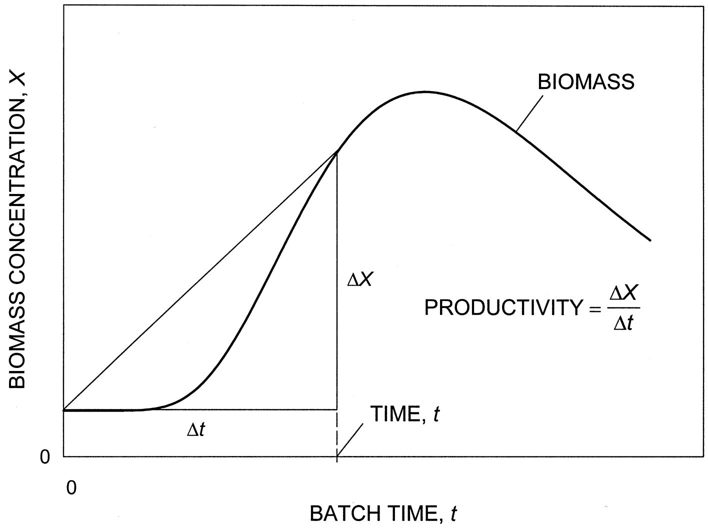


FIGURE 31 Biomass productivity in batch culture.

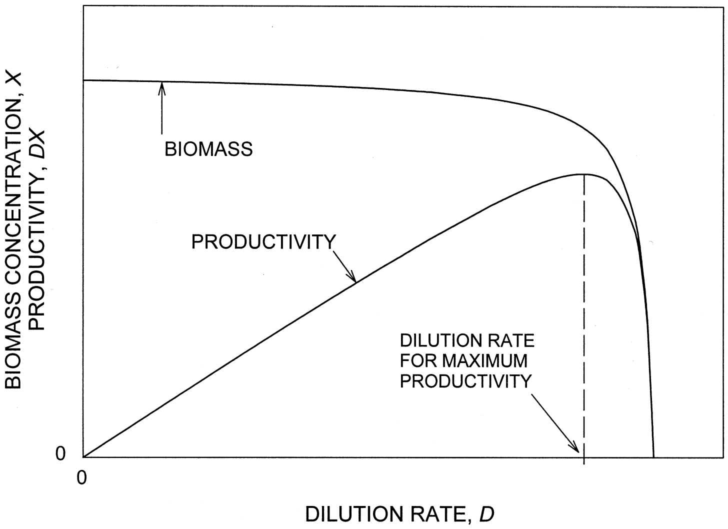


FIGURE 32 Variation of the steady sate biomass concentration and the productivity with dilution rate in a well-mixed continuous flow bioreactor.

washout of cells from the bioreactor. In practice, the dilution rate must remain quite a bit less than the value needed for optimal productivity (Fig. 32), or cells may be washed out because of an inadvertent slight increase in the dilution rate. If viable biomass from the outflow of a continuous flow well-mixed bioreactor is recycled to the reactor, as in Fig. 23b, then the reactor may be operated at a dilution rate greater than the maximum specific growth rate. Such a bioreactor provides a greater biomass productivity compared to one that does not recycle the biomass and the biomass concentration in the reactor outflow stream is also greater.

3. Enzyme Kinetics

Instead of viable cells, a bioreactor may use nonviable cells and isolated enzymes as the biocatalyst. The reaction in such a bioreactor may obey Michaelis-Menten kinetics but other kinetic patterns are also observed. For a reaction that obeys Michaelis-Menten kinetics, the rate of reaction (i.e., the rate of consumption of the substrate, $-dS/dt$) depends on the concentrations of the substrate S and the enzyme e, as follows:

$$-\frac{dS}{dt} = \frac{k_r e S}{K_m + S}, \tag{9}$$

where k_r is the rate constant and K_m is known as the Michaelis-Menten constant. The dynamics of a bioreactor are often analyzed in terms of the *conversion* C_s of the substrate, where C_s is the fraction of the original substrate transformed to a product, i.e.,

$$C_s = \frac{S_o - S}{S_o}. \tag{10}$$

In Eq. (10), S_o is the initial concentration of the substrate in the reactor and S is the concentration at time t.

For Michaelis-Menten kinetics in a well-mixed batch bioreactor, the conversion of the substrate at any time t is governed by the relationship:

$$\frac{k_r E t}{V_L} = S_o C_s - K_m \ln(1 - C_s), \tag{11}$$

where E is the total amount of enzyme in a bioreactor of volume V_L. When the reaction is carried out in a continuous flow well-mixed bioreactor, the expression for the conversion is as follows:

$$\frac{k_r E}{F} = S_o C_s + \frac{K_m C_s}{1 - C_s}, \tag{12}$$

where F is the volume flow rate of the feed. Similarly, in a packed bed bioreactor, the expression for the conversion is the following:

$$\frac{k_r E}{F} = S_o C_s + K_m \ln(1 - C_s). \tag{13}$$

Because F is the volume processed in time t in a continuous flow bioreactor and V_L is the corresponding volume in a batch reactor, a comparison of Eqs. (11) and (13) shows that batch and plug flow (i.e., packed bed) bioreactors containing the same amount of enzyme will achieve equal conversions in a given time. This is a general conclusion, irrespective of the reaction kinetics. A continuos flow packed bed enzyme bioreactor may be advantageous relative to batch reactor, as the unproductive time for batch preparation could be eliminated in the continuous flow unit. However, the batch reactor may have other important advantages such as the ease of pH control in a well-mixed device.

When the substrate concentration S is much greater than K_m, Eqs. (12) and (13) reduce to the same form. In this case, the continuous flow stirred reactor and the plug flow device achieve similar conversion values in a given time. In contrast, when $S \ll K_m$, the reaction rate becomes first order in the substrate concentration (see Eq. (9)), and the plug flow reactor provides higher conversion values in comparison with the well-mixed continuous flow device. In the latter bioreactor, all the enzyme would be exposed to the same low concentration of the substrate which is not useful except when the reaction is inhibited by the substrate.

IV. CONCLUDING REMARKS

A bioreactor is an indispensable part of any bioprocess irrespective of whether the process degrades pollutants or produces substances such as foods, feeds, chemicals and pharmaceuticals, and tissues and organs for use in biomedicine. The variety of bioprocesses is tremendous and many different designs of bioreactors have been developed to meet the different needs. In all cases, the bioreactor must provide the environmental conditions necessary for the culture. The specific demands are often conflicting and achieving optimal performance requires attaining the proper balance among the different requirements. Success of a bioprocess depends critically on good design and operation of the bioreactor.

SEE ALSO THE FOLLOWING ARTICLES

BIOENERGETICS • BIOMASS, BIOENGINEERING OF • ENZYME MECHANISMS • MAMMALIAN CELL CULTURE • METABOLIC ENGINEERING • PHOTOCHEMISTRY, MOLECULAR • SEPARATION AND PURIFICATION OF BIOCHEMICALS

BIBLIOGRAPHY

Atkinson, B. (1974). "Biological Reactors," Pion Press, London.

Bailey, J. E., and Ollis, D. F. (1986). "Biochemical Engineering Fundamentals," 2nd Ed., McGraw-Hill, New York.

Chisti, Y. (1989). "Airlift Bioreactors," Elsevier, London.

Chisti, Y. (1999). Solid substrate fermentations, enzyme production, food enrichment. *In* "Encyclopedia of Bioprocess Technology: Fermentation, Biocatalysis, and Bioseparation," (M. C. Flickinger and S. W. Drew, eds.), Vol. 5, pp. 2446–2462, Wiley, New York.

Chisti, Y. (1999). Modern systems of plant cleaning. *In* "Encyclopedia of Food Microbiology" (R. Robinson, C. Batt, and P. Patel, eds.), pp. 1806–1815, Academic Press, London.

Chisti, Y., and Moo-Young, M. (1999). Fermentation technology, bioprocessing, scale-up and manufacture. *In* "Biotechnology: The Science and the Business" (V. Moses, R. E. Cape, and D. G. Springham, eds.), 2nd ed., pp. 177–222, Harwood Academic Publishers, New York.

Deckwer, W. D. (1992). "Bubble Column Reactors," Wiley, New York.

Doran, P. M. (1999). "Design of mixing systems for plant cell suspensions in stirred reactors," *Biotechnol. Prog.* **15,** 319–335.

Moser, A. (1981). "Bioprocess Technology," Springer-Verlag, New York.

Nienow, A. W. (1998). "Hydrodynamics of stirred bioreactors," *Appl. Mech. Rev.* **51,** 3–32.

Tredici, M. R. (1999). Bioreactors, photo. *In* "Encyclopedia of Bioprocess Technology: Fermentation, Biocatalysis, and Bioseparation" (M. C. Flickinger and S. W. Drew, eds.), Vol. 1, pp. 395–419, Wiley, New York.

Van't Riet, K., and Tramper, J. (1991). "Basic Bioreactor Design," Dekker, New York.

Varley, J., and Birch, J. (1999). "Reactor design for large scale suspension animal cell culture," *Cytotechnology* **29,** 177–205.

Willaert, R. G., Baron, G. V., and De Backer, L. (eds.) (1996). "Immobilised Living Cell Systems: Modelling and Experimental Methods," Wiley, Chichester.

Fiber-Optic Chemical Sensors

David R. Walt
Israel Biran
Tarun K. Mandal
Tufts University

I. Introduction
II. Fundamental Principles of Fiber-Optic Chemical Sensors
III. Sensing Schemes for Fiber-Optic Chemical Sensors
IV. Applications of Fiber-Optic Chemical Sensors
V. Recent Developments
VI. Conclusions

GLOSSARY

Absorption Process by which electromagnetic energy is transferred to an atom or molecule. Different molecules absorb light at different discrete wavelengths and therefore each molecule has a characteristic absorption spectrum.

Antibody Biologically derived molecule able specifically to recognize and bind to another molecule (antigen).

Biosensor Sensor in which the sensing material is of biological origin (e.g., enzyme, antibody, cell, DNA).

Extrinsic sensor Sensor based on indirect analysis using an indicator (compound that changes optical properties by reaction with an analyte).

Fluorescence Emission of light at a longer wavelength by molecules or other chemical species after excitation at a shorter wavelength.

Imaging fiber Bundle of coherently ordered individual optical fibers through which an image can be transmitted from one side to the other.

Immobilization Attachment of chemical species to surfaces, polymers, or other insoluble substrates. For example, with optical sensors, indicators are attached to the optical fiber surface.

Indicator Molecule that undergoes changes in optical properties on interaction with chemical species.

Intrinsic sensor Sensor based on the direct spectroscopic detection of a molecule of interest.

Optical fiber Transparent filament made out of glass or plastic through which light can propagate by total internal reflection.

Sensing element Combination of sensing materials (indicators and/or biological compounds) and a transducer (optical fiber).

Sensing material Material (indicator, dye, polymer,

biological molecule) that is immobilized to the optical fiber surface and can change its physicochemical properties upon interaction with an analyte.

Total internal reflection Reflection of light at the interface between two materials of different refractive index. In optical fibers, the core refractive index is higher than the clad refractive index. When light is introduced into the core, it is reflected at the core–clad interface.

A CHEMICAL SENSOR is an analytical device that can measure the concentration of a specific chemical or a group of chemicals in a sample of interest. The basic structure of a chemical sensor includes (a) a sensing material that selectively interacts with the analyte and (b) a transducer (e.g., electrochemical, optical, thermal, or mass) that can transform this interaction into a measurable signal. This signal should be proportional to the magnitude of the changes in the physicochemical properties associated with the interaction between the sensing material and the analyte. Ideally, chemical sensors should operate in a continuous and reversible manner.

Fiber-optic chemical sensors are analytical devices incorporating optical fibers as part as their optical transducing system. Optical fibers are small and flexible "wires" made out of glass or plastic that can transmit light signals, with minimal loss, for long distances. The signals generated by the sensing materials, which are usually immobilized to the fiber surface, are transmitted through the optical fibers and can be measured by using different optical methods such as absorption, fluorescence, and Raman spectroscopy. Fiber-optic chemical sensors can be used for remote analytical measurements in applications including clinical, environmental, and industrial process monitoring. Several fiber-optic chemical sensors are commercially available and it is expected that recent developments in optical technologies and the research efforts to use these technologies for fiber-optic sensor development will lead to a number of commercially available fiber-optic chemical sensors for various applications.

I. INTRODUCTION

Fiber-optic chemical sensors are composed of a sensing material and a transducer. The transducer converts the recognition and sensing events obtained by the sensing materials into a response such as an optical signal. Optical measurements can provide rapid, sensitive, and nondestructive analysis of many important compounds. In fiber-optic chemical sensors, optical fibers are used to transmit the optical signal to the measurement device, enabling a remote detection of the analyte in the sample. The use of optical fibers for sensing applications was first suggested in the mid-1960s. Since then, various fiber-optic chemical sensor types have been developed for detecting numerous analytes. In the last decade, due to the rapidly growing use of fiber optics for telecommunication applications, new fiber-optic technologies have been developed resulting in high-quality and inexpensive optical fibers that can be used for chemical sensing applications. Optical fibers are remarkably strong, flexible, and durable. These features and their nonelectrical nature make them highly suitable for different industrial and environmental applications where safe sensing in hazardous and harsh environments is needed (e.g., monitoring of chemicals in nuclear plants or toxic gases in petrochemical plants). Fiber-optic chemical sensors are also widely used in the clinical field since their small dimensions allow them to be used for *in vivo* sensing thereby eliminating the need to procure samples. Furthermore, as described later in Section IV, fiber-optic chemical sensors can be incorporated into optical fiber bundles used for *in vivo* imaging (endoscopes) to provide both analytical information and imaging capabilities. One significant advantage of employing optical fibers is that multiple optical signals can be transmitted and measured simultaneously, thereby offering multiplexing capabilities.

Optical fibers and the instrumentation used in fiber-optic chemical sensors are described in Section II. We describe how different optical phenomena, generated by different sensing mechanisms, can be applied to optical fibers to measure analytical signals. These sensing mechanisms are described in Section III. Section IV reviews several fiber-optic chemical sensor analytical applications in the clinical, industrial, and environmental fields and Section V reviews recent developments in fiber-optic chemical sensors.

II. FUNDAMENTAL PRINCIPLES OF FIBER-OPTIC CHEMICAL SENSORS

The basic design of a fiber-optic chemical sensor system is shown in Fig. 1. The fiber-optic chemical sensor's main components are (a) a light source, (b) optical fibers to both transmit the light and act as the substrate for (c) the sensing material, and (d) a detector to measure the output light signal. Usually computers or microprocessors control the fiber-optic chemical sensor instrumentation and are employed to analyze the output signals.

In this section, optical fibers and their basic characteristics are described. Fiber-optic chemical sensor instrumentation and the optical phenomena employed are also described.

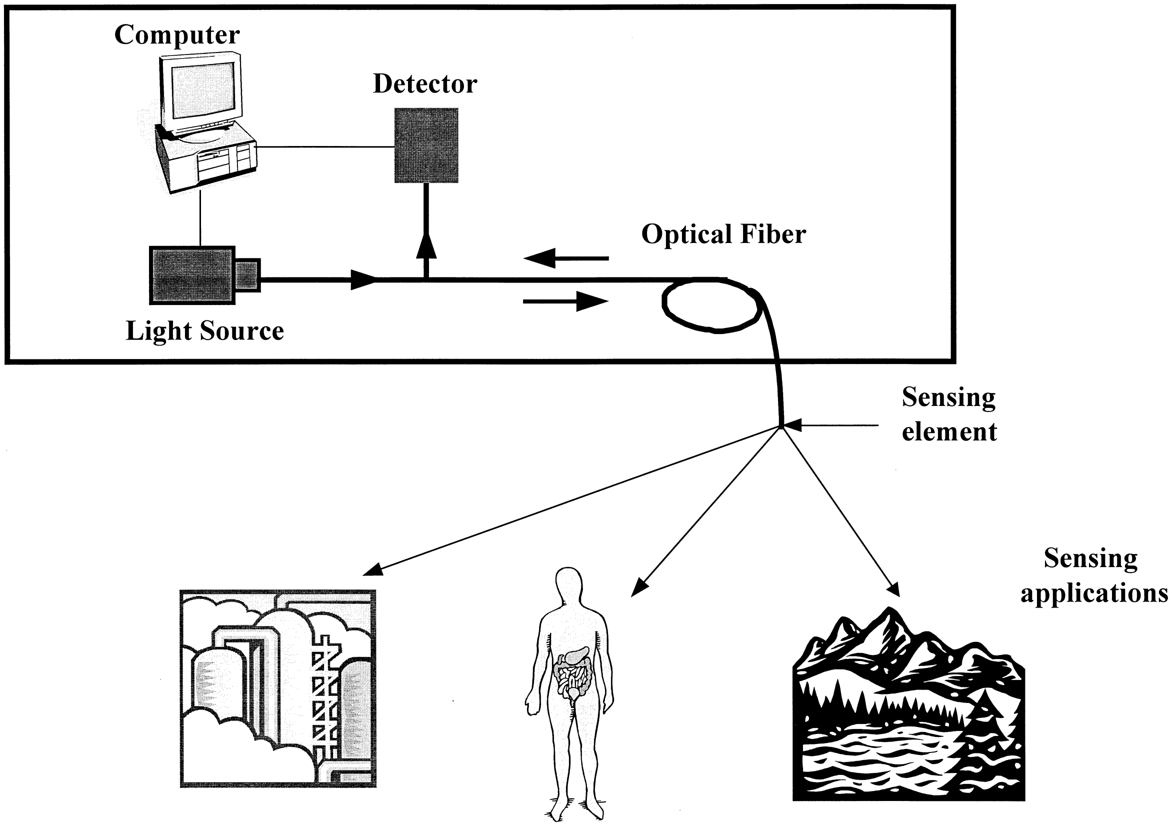


FIGURE 1 Schematic diagram of a fiber-optic chemical sensor system with examples of environmental, clinical, and industrial applications.

A. Optical Fibers

1. Basic Characteristics

Optical fibers are waveguides made out of glass or plastic, through which light can be transmitted. Optical fibers transmit light very efficiently, which is why they are so useful for many applications. The light transmission through the fiber is based on the phenomenon of total internal reflection (TIR). Optical fibers consist of a core with a refractive index n_1 surrounded by a cladding with a lower refractive index n_2 (Fig. 2). The difference between the refractive indices enables the core–clad interface to effectively act as a mirror such that a series of internal reflections transmits the light from one end of the fiber to the other as shown in Fig. 3a. Several principles related to the light transmission through the optical fiber are significant for fiber-optic chemical sensor function and design:

1. *The Critical Angle.* If light strikes the cladding at an angle greater than the critical angle φ_c, the light is totally internally reflected at the core–clad interface (Fig. 3a). If light strikes the cladding at an angle less than the critical angle, as shown in Fig. 3b, it is partly reflected and partly refracted. The critical angle is defined by the ratio between the clad and the core refractive indices,

$$\sin \varphi_c = n_2/n_1. \tag{1}$$

2. *Acceptance Cone.* In order to get high light transmission, light should propagate through the fiber by a series of total internal reflections. This transmission is achieved if the angles of the light entering the fiber are within the acceptance cone as shown in Fig. 3c. The acceptance cone

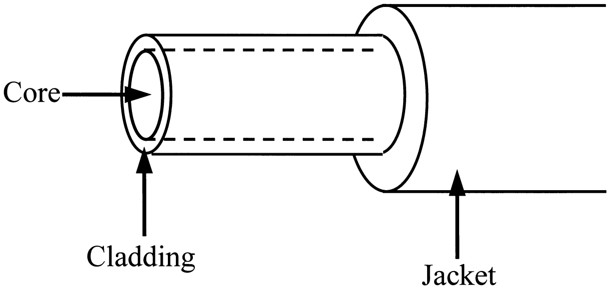


FIGURE 2 Schematic diagram of an optical fiber showing core and clad structure.

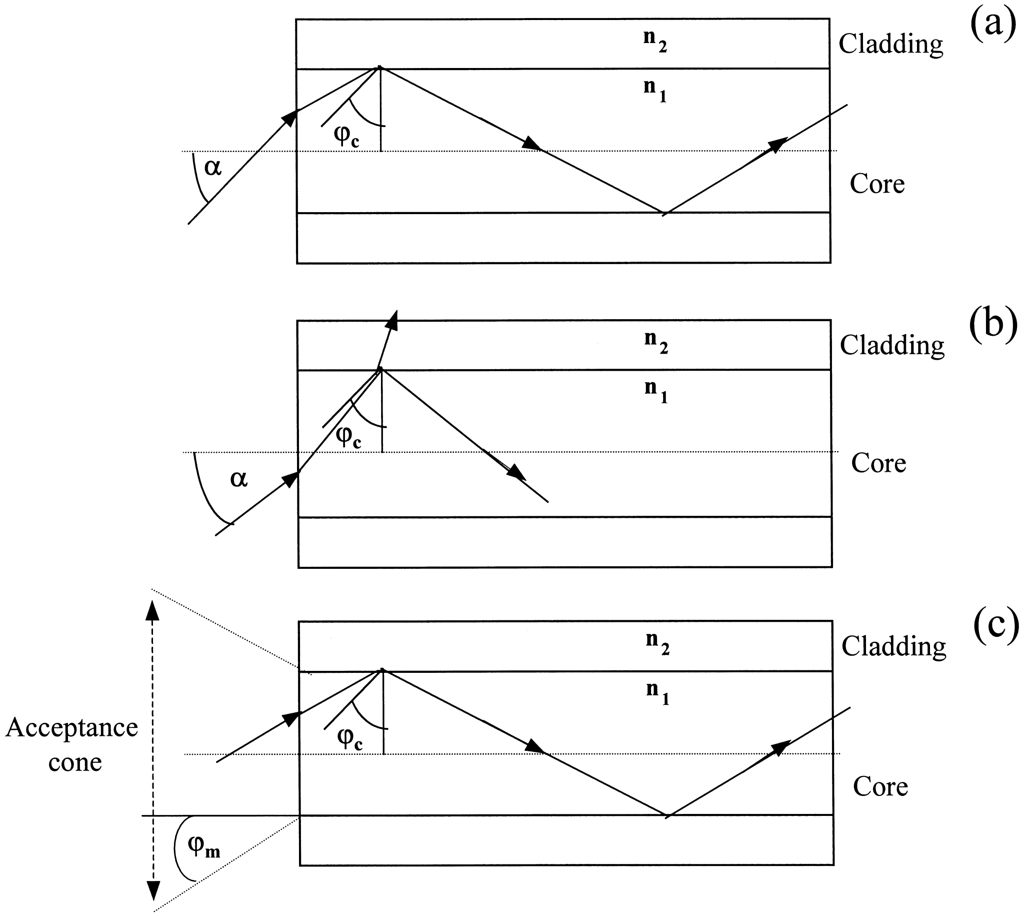


FIGURE 3 Propagation of light through the optical fiber occurs when the total internal reflection condition exists at the interface between the core (index of refraction n_1) and cladding (n_2) such that $n_1 > n_2$. (a) Light entering the fiber is totally internally reflected (TIR) if the light angle is greater than the critical angle φ_c. (b) Light will be partly reflected and partly refracted if the light angle is less then the critical angle φ_c. (c) Light will propagate in TIR only when the entering light angle is within the acceptance angle.

size depends on the refractive indices of the core and the clad and also on the refractive index of air n_0,

$$\sin\varphi_m = \frac{\sqrt{(n_1^2 - n_2^2)}}{n_0}. \quad (2)$$

3. *Numerical Aperture.* The acceptance cone's width determines the efficiency of light collection of the fiber and can also be described in terms of the numerical aperture (NA),

$$\mathrm{NA} = n_0 \sin\varphi_m. \quad (3)$$

A high NA indicates a wide acceptance cone and better light-gathering capabilities of the fiber. A typical NA value for a high-quality glass fiber is 0.55 but fiber NAs as high as 0.66 or as low as 0.22 have been used for sensing.

2. Optical Fiber Types

Optical fibers are usually made of glass in a process that involves heating a glass preform and then drawing the fiber from the melted hot glass. In most fiber-optic chemical sensors, glass fibers are used to transmit light in the visible and near-infrared regions of the optical spectrum ($400 < \lambda < 700$ nm). In some cases, when the light employed is in the UV region, quartz (pure silica) fibers are used as the core material and doped silica (with lower refractive index) is employed as the cladding material. Optical fibers made of silver halide or chalcogenides are used to transmit light in the infrared region of the spectrum ($\lambda > 700$ nm). Plastic fibers are also used for fiber-optic chemical sensors; these fibers are very flexible and cheap but their optical characteristics are inferior to those of glass fibers and their heat tolerance is lower.

3. Optical Fiber Configurations

Optical fibers are produced in many different configurations, formats, and sizes. For some fiber-optic chemical sensor applications, single optical fibers with diameters ranging from 50 to 500 μm are employed. For

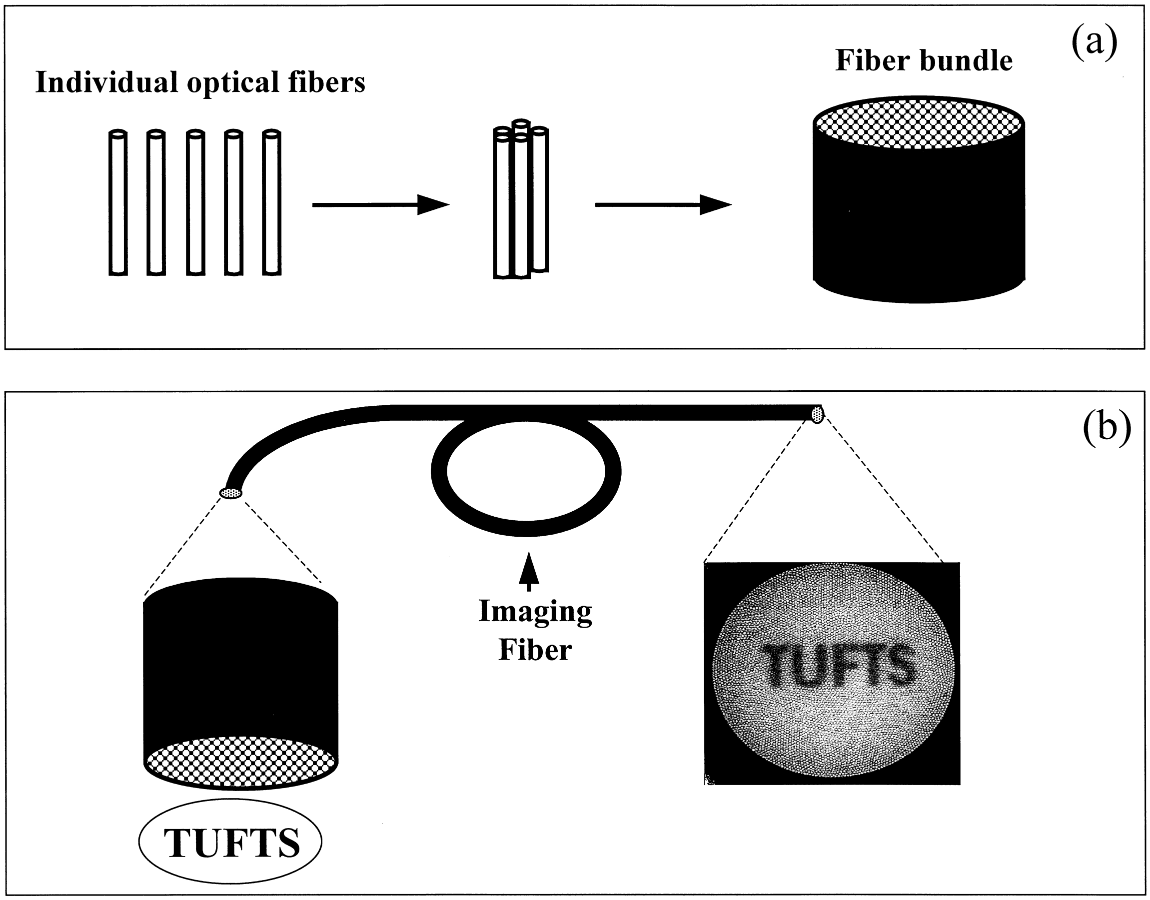


FIGURE 4 Optical fiber bundle fabrication and use for imaging. (a) Fiber bundles are constructed from thousands of individual single fibers that are fused together. (b) Coherent bundles can be used for imaging.

other fiber-optic chemical sensor applications, fiber-optic bundles comprising thousands of identical single fibers (each with a diameter of a few micrometers) are employed, as shown in Fig. 4a. The fibers can be bundled in a coherent or random fashion and used to enhance light transmission. In coherent fiber bundles, the position of each fiber on one end is identical to its position on the other end. These coherent fiber bundles are mostly used for imaging applications as shown in Fig. 4b. The coherent fibers allow each individual fiber to transmit light in an addressable manner from one end of the bundle to the other. In some recent applications, these bundles are used for imaging and sensing simultaneously.

B. Fiber-Optic Chemical Sensor Design and Instrumentation

Each fiber-optic chemical sensor system design depends intimately on the nature of the analytes detected, the particular conditions at the measurement site, and the specific requirements for the application, for example, the use of nontoxic agents and biocompatible materials for clinical applications. There are two basic configurations for fiber-optic chemical sensors: (a) A single fiber is used to transmit the light from the light source to the sample region and back to the detector, as shown in Fig. 1. (b) Multiple fibers are used, one to transmit the light to the sample region and the rest to transmit light from the sample region to the detector, as shown in Figs. 5a and 5b. For the second configuration, the most common format is a bifurcated fiber. Bifurcated fibers are fabricated by fusing two fibers together on one end, leaving the other ends free. The fused side is used for the sensing and the other ends of the fiber are connected to the light source and to the detector, as shown in Fig. 5c.

1. Light Sources

The light sources used for fiber-optic chemical sensors should provide sufficient light intensity within the sensor

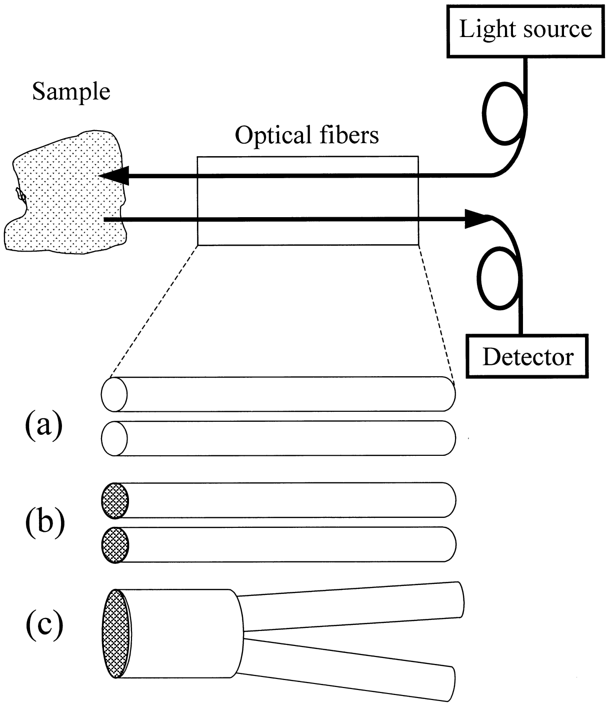


FIGURE 5 Design principle of fiber-optic chemical sensors. (a) Two single fibers. (b) Two fiber bundles. (c) Bifurcated fiber.

wavelength operating range. In addition, the light output should be stable over long time periods since light fluctuations may add noise to the measurement and reduce the sensor sensitivity. Tungsten–halogen lamps are often used as a light source to provide high-intensity light and a continuous emission spectrum over wavelengths ranging from 300 to 2700 nm. This lamp is very useful because of its small size and high stability, and the potential to use it in portable devices. Light-emitting diodes (LEDs) are also used as light sources for fiber-optic chemical sensors because they offer spectral consistency and small size; however, they are limited in their light intensity. Lasers are an important light source for fiber-optic chemical sensors since they supply very intense monochromatic light. Lasers are used for fluorescence-based fiber-optic chemical sensors in which fluorophores with low quantum yields are used.

2. Optical Signal Detectors

In most fiber-optic chemical sensor systems, the light transmitted from the sensing element (output light) is measured by using photon detection devices, which absorb photons and convert them into electrical signals. Several photon detectors are available. The simplest and cheapest detector is the silicon photodiode. In this solid-state device, photon detection is based on *p–n* semiconductor junctions. When photons strike the *p* junction, electrons move into the conducting band and act as charge carriers. This process generates a current that is proportional to the incident light intensity. A series of photodiodes can be arranged in a linear array and used to detect the entire spectrum simultaneously. These devices offer advantages such as small size, wide spectral range, and low power requirement; however, photodiode sensitivity is relatively low. Photomultiplier tubes (PMT) are more sensitive than photodiodes and are used for low-light-level detection. Photons that strike the photocathode cause the ejection of photoelectrons that are attracted to the photomultiplier electrodes (also called dynodes). The dynodes are arranged such that each dynode is held at a higher potential than the previous one. When a high-energy photoelectron strikes the first dynode, it causes the ejection of several photoelectrons from the dynode surface. These cascade processes continue as more photoelectrons strike each subsequent dynode, resulting in a high amplification of the initial photon. In addition to their high sensitivity, PMTs have high signal-to-noise ratios and respond very fast. PMT detectors, however, are temperature sensitive and can be irreversibly damaged if exposed to high light levels.

In some fiber-optic chemical sensors, the optical signal measurement involves image acquisition and analysis. Charge-coupled device (CCD) chips are typically used to capture the output image. The CCD is a two-dimensional integrated-circuit chip array made up of several layers of silicon and an array of gate electrodes. These electrodes, often referred as metal-oxide semiconductors (MOS), are individually connected to an external voltage source. When photons strike the chip's light-sensitive silicon layer, they create electrical charges that are detected by the gate electrodes. This detection process is called charge coupling and involves the transfer of voltage between the two gate electrodes. Since each pair of gate electrodes can detect the localized photon strikes, intensity and spatial position of the image are simultaneously measured. CCD chips offer additional advantages including their small size (a few square millimeters to a few square centimeters), high sensitivity, and stability. The limitations of CCD chips are the requirement for low-temperature operation and their relatively high cost.

C. Optical Phenomena Employed for Sensing in Fiber-Optic Chemical Sensors

Chemical sensing via optical fibers is performed by monitoring the changes in selected optical properties. Interactions of light with the analytes are measured either directly or by using indicators. The light is launched from a light source into the fiber and guided to the measurement site. The absorbed, reflected, scattered, or emitted light

can be measured using several different optical phenomena. These phenomena transduce the interactions of light with the sensing materials into an (ideally) quantitative signal that can be correlated to the analyte identities and concentrations.

1. Absorption

Absorption is based on the light intensity changes due to modulation by a substance in the sample. Light absorption is a process in which electromagnetic energy is transferred to an atom or a molecule. This energy promotes the transition of the molecule from the ground energy state to a higher energy excited state. The resulting energy is dissipated nonradiatively (i.e., thermally) to the medium when the excited state relaxes to the ground state. Each molecule (analyte) can be excited at a single wavelength or several wavelengths, which furnishes a unique absorption spectrum characteristic of the molecule. The absorbance changes are related to the analyte concentration [C] via the Beer–Lambert relationship:

$$A = \log(I_0/I) = \varepsilon \cdot [\mathrm{C}] \cdot l, \qquad (4)$$

where A is the optical absorbance, I_0 and I are the intensities of transmitted light in the absence and presence of the absorbing species, respectively, l is the effective path length, and ε is the molar absorption coefficient. In practice, optical fibers are connected to a spectrophotometer and the measured changes correlate the analyte concentration to the absorption at a given wavelength.

2. Fluorescence

When fluorescent molecules are excited at a specific wavelength, the molecule reemits radiation at a lower energy, i.e., a longer wavelength. The absorption of the excitation light shifts the molecule's energy from the ground state to a higher energy state. The molecule emits fluorescent light when it returns to the ground state. The distinct ranges of wavelengths over which the molecule is excited and emits are well defined and simple to detect, as shown in a typical spectrum of a fluorescent molecule in Fig. 6.

Concentrations of the fluorescent analytes are measured by transmitting an excitation light through the optical fiber and measuring the light emission intensity using a detector. A nonfluorescent analyte can be measured indirectly if its interaction with an indicator molecule changes the indicator emission intensity (see Section III.B).

A decrease in fluorescent intensity due to fluorescence quenching can also be used for sensing. In this case, the analyte's interaction with a fluorescent molecule causes a decrease in fluorescence (quenching). The magnitude of the fluorescence decrease is related to the analyte concentration.

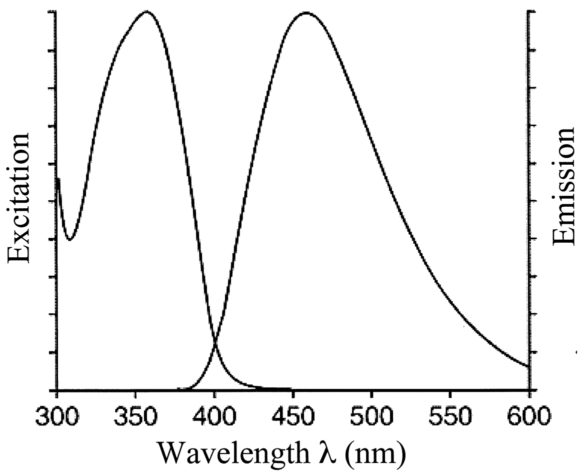


FIGURE 6 Typical fluorescence spectrum showing the strokes shift at longer wavelengths from the excitation spectrum.

3. Time-Resolved Fluorescence Spectroscopy

This method is based on the excited-state lifetime. The light intensity emitted from molecules excited by a short pulse of light decays exponentially with time. This decay pattern is unique for each molecule and can be used for analytical purposes. Alternatively, a phase shift method can be employed to measure the fluorescence lifetime. A sinusoidally varying excitation light source is used and the phase shift between the excitation waveform and the emission waveform can be used to detect the analytical signal.

4. Fluorescence Energy Transfer

This phenomenon occurs when two distinct fluorophores are present. If the emission spectrum of one fluorophore overlaps with the excitation spectrum of a second fluorophore and the two fluorophores are in sufficient proximity (<100 Å), then the excited fluorophore (donor) can transfer energy nonradiatively to the second fluorophore (acceptor). This transfer results in an increase in light emission by the acceptor and a decrease in light emission from the donor. When an energy transfer pair of fluorophores is used to label two molecules that can interact (antibody–antigen, enzyme–substrate), they can be use for sensing in fiber-optic chemical sensors.

5. Raman Spectroscopy

In Raman spectroscopy, light is scattered from the molecule in different directions and is shifted to both higher and lower frequencies. The shift in magnitude is equal to the characteristic vibration frequencies of the molecule, resulting in a unique spectrum for each molecule. Optical fibers are used as light guides for Raman spectroscopy because the optimum wavelengths for the

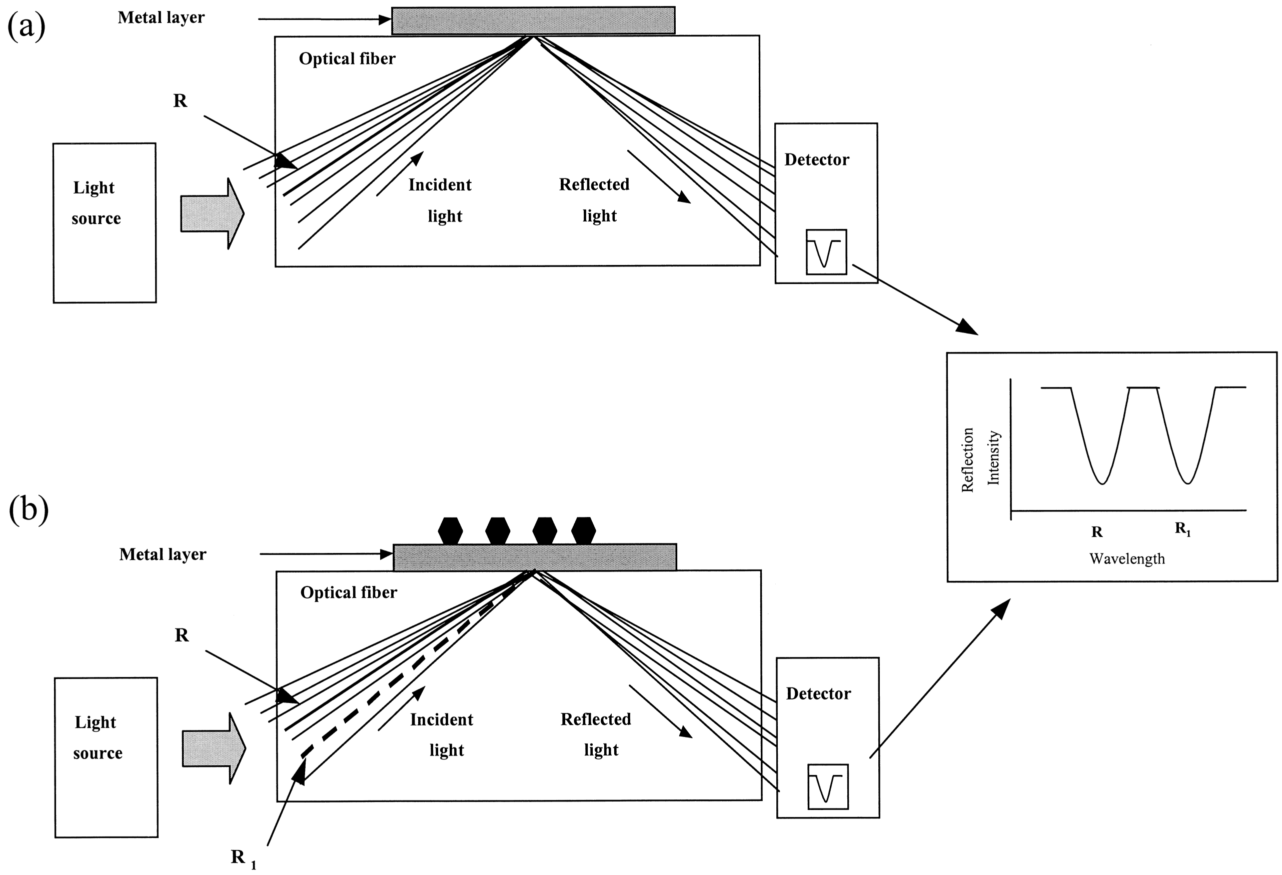


FIGURE 7 Schematic principle of surface plasmon resonance (SPR) sensing. (a) Light is absorbed only at the resonant angle R. (b) Analyte binding shifts the light resonant angle to angle R_1.

analysis of many different analytes are in the visible or near-IR range. The limitation to the Raman method is the low intensity of the scattered light. A significant enhancement of the scattered light signal (10^6) is observed when the analyte molecules are adsorbed on metal surfaces. In this method, called surface-enhanced Raman scattering (SERS), two optical fibers are typically used. One fiber is used to transmit the excitation light to the sensing area (glass slide covered with a metal layer, usually silver) and the second fiber is used to collect the scattered radiation.

6. Surface Plasmon Resonance (SPR)

Very small changes in the refractive index on the sensor surface can be measured using SPR, including changes caused by the binding of molecules. SPR occurs at the surface of a thin metal layer placed on a glass slide or other dielectric material such as an optical fiber. Incident light with a specific angle is resonantly coupled to the surface plasmons (oscillating electrons at the edges of the metal) in the metal layer because of their matched frequencies. Energy is absorbed in this resonance, resulting in a sharp decrease in the reflected light (Fig. 7). This phenomenon is used for analytical applications since any changes at the metal layer surface, i.e., binding of analyte, cause a shift in the resonant angle. The magnitude of the shift from the initial resonant angle before the binding of the analyte is proportional to the concentration of the bound analyte.

7. Evanescent Wave Spectroscopy

When light propagates by total internal reflection (TIR) through an optical fiber or a planar optical waveguide, a small portion of the electromagnetic wave penetrates the region of lower refractive index, thereby generating an evanescent field that exponentially decays as a function of the distance from the fiber or waveguide surface (Fig. 8a). When a small part of the fiber cladding is removed, the evanescent field around the exposed core can be used for sensing at the core–air or core–liquid interface. Binding of molecules to the exposed region may induce several measurable changes in the properties of the light transmitted

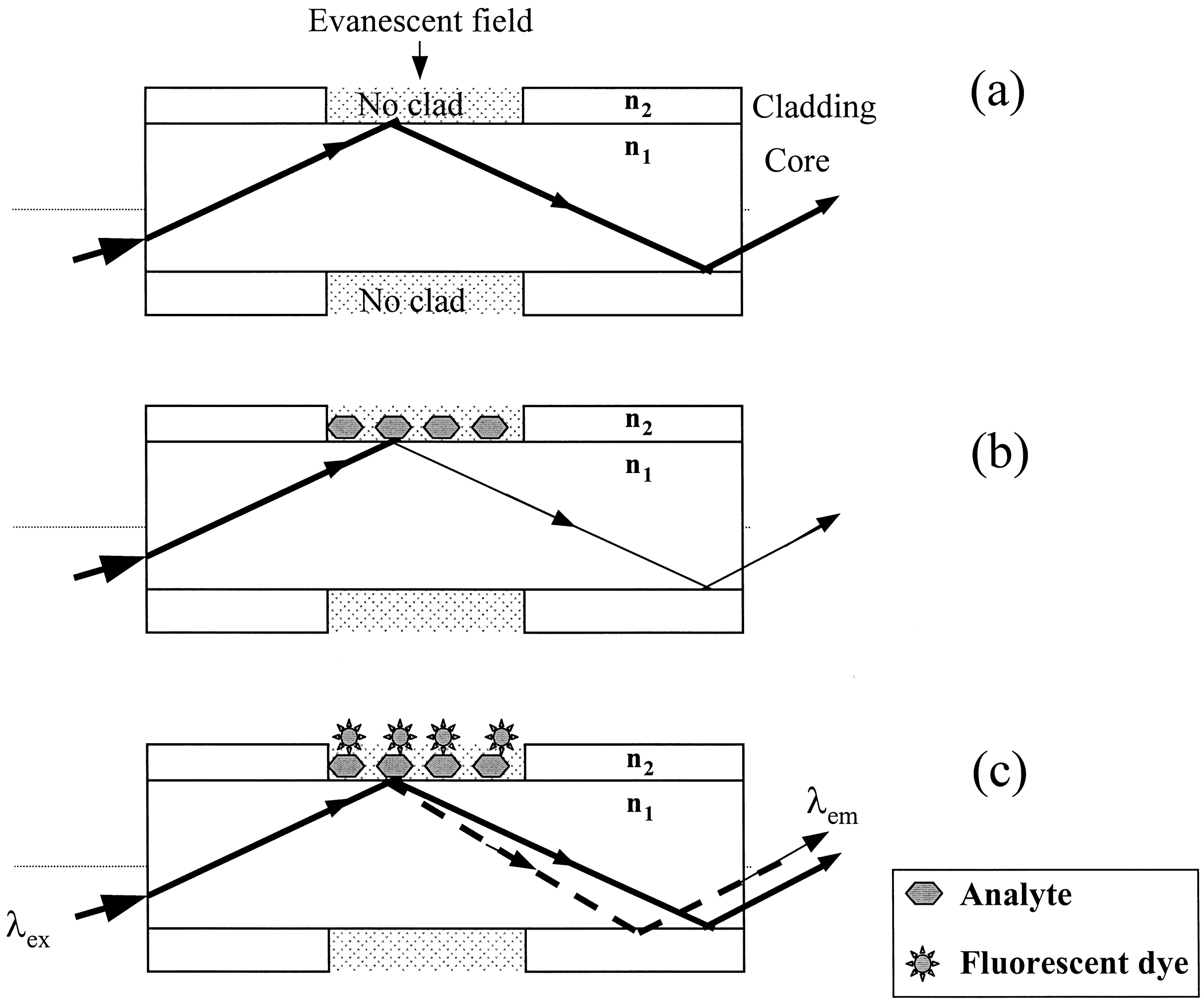


FIGURE 8 Schematic principle of evanescent wave field sensing. (a) The evanescent field is formed when a small portion of the cladding is removed. (b) The bound analytes absorb some of the light propagating through the fiber. (c) Excitation light (λ_{ex}) transmitted through the fiber excites the bound fluorescent molecules and emitted light (λ_{em}) is measured at the fiber output.

through the fiber. In one example, the target molecules absorb some of the light and reduce the transmitted light intensity, as shown in Fig. 8b. Alternatively, the binding of fluorescent molecules can be determined by employing light that corresponds to the molecule's excitation wavelength and measuring the emitted light, as shown in Fig. 8c.

III. SENSING SCHEMES FOR FIBER-OPTIC CHEMICAL SENSORS

Fiber-optic chemical sensors can be divided into two categories based on their structure: (a) Intrinsic sensors are based on the analyte's intrinsic optical properties, and (b) extrinsic sensors are based on sensing materials (chemical or biological) immobilized to the fiber surface, with the optical fibers used only as a transmission pathway between the sensing materials and a remote measurement device. The basic concepts of these two sensing schemes are shown in Fig. 9. Optical fibers allow measurements to be made directly in the sample. Intrinsic sensors usually have a simple structure and a fast response time, but they are not as selective as extrinsic sensors because many groups of compounds exhibit similar optical properties when measured directly. Extrinsic sensors provide an additional level of selectivity. In this section, we describe both sensor types and their use for fiber-optic chemical sensors.

A. Intrinsic Sensing Mechanism—Direct Spectroscopy

Many different spectroscopic techniques are used for direct spectroscopic measurements. These techniques

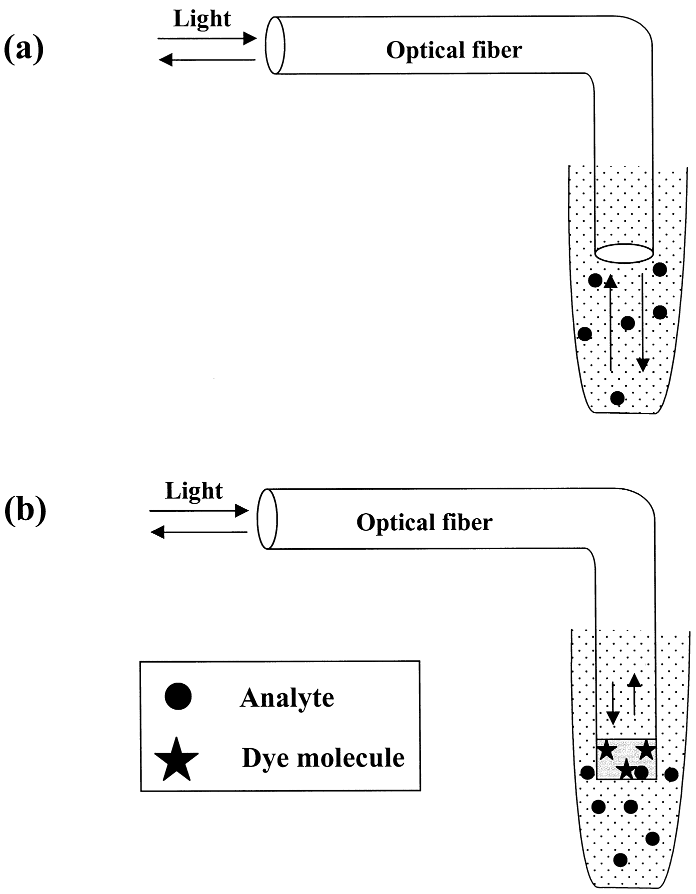


FIGURE 9 Two basic fiber-optic chemical sensing schemes: (a) Intrinsic, (b) extrinsic.

include transmission or absorption, attenuated total reflection, photoacoustic, fluorescence, light scattering, infrared, and Raman spectroscopy. In this section, we describe how some of these techniques are employed for use with optical fibers to measure the concentration of chemical species either in the liquid phase or the gas phase. Generally, an optical fiber is coupled with an instrument such as a spectrophotometer or fluorimeter. The particular optical fibers used are determined by the specific requirements of each type of spectroscopic technique.

1. Liquid-Phase Sensing

Transmission spectroscopy has found wide application in chemical, biological, and environmental monitoring because of its intrinsic safety and ease of application. This method generally is used to monitor a single analyte in a harsh analytical medium. Transmission spectroscopy is also used for simultaneous detection of several species at trace levels in a complex medium. Light is focused into the fiber, passes through the medium, and then is coupled through the return fiber into the input slit of the spectrometer. The light losses due to the small core diameter and acceptance angle of the optical fiber are reduced by attaching a convergent lens at the remote sensor head (Fig. 10). This lens causes the light from the illumination fiber to pass through the sensing region and refocuses the transmitted light into the return fiber. Using this simple technique with an optical fiber allows one to measure the absorbance of the analyte directly. For example, the salt concentration in an electroplating bath can be easily measured by monitoring the solution absorbance. This technique is also used for *in vivo* analysis of components in human blood (see Section IV.A).

Optical fibers are commonly used for remote monitoring of fluorescent analytes. Fiber-optic chemical sensors can provide both qualitative and quantitative information about the analyte under consideration. Since each analyte has different fluorescent properties, selective measurements can be performed by choosing the correct excitation and emission wavelengths. The fluorescence bands are usually quite broad and the bands for a class of compounds overlap, so it may be difficult to distinguish among them. Fluorescence lifetime measurements are sometimes

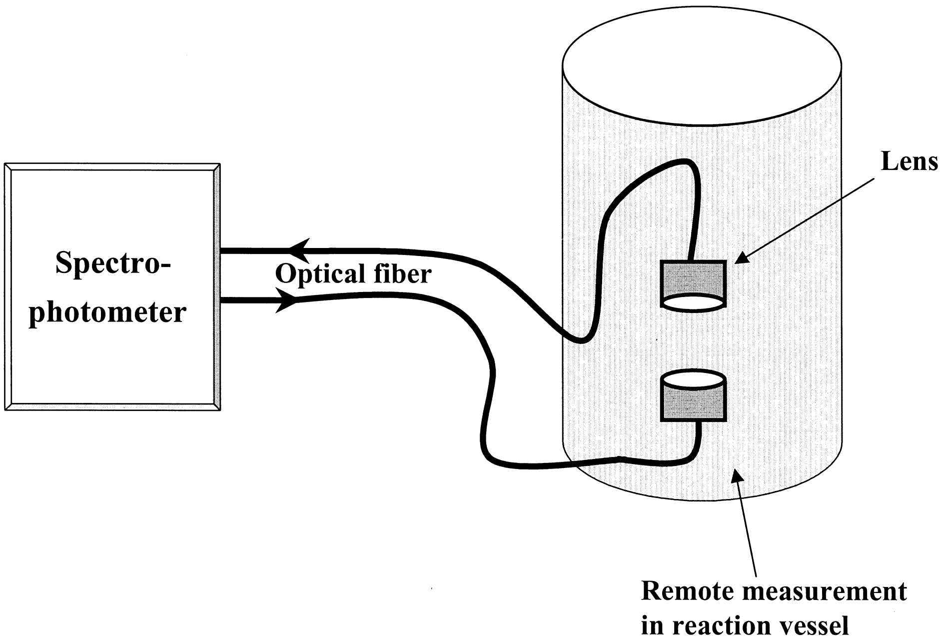


FIGURE 10 A typical arrangement for remote measurement of sample absorption using optical fibers.

used in such cases, as they provide an additional parameter for distinguishing such compounds. For example, phenol, toluene, and xylene are single-ring aromatic compounds with similar emission peaks. It is not possible to distinguish between them by simply measuring their fluorescence spectra. It is possible, however, to distinguish between them by measuring their fluorescence decay times. Also, low concentrations of dissolved polyaromatic hydrocarbons, such as benzene, naphthalene, and pyrene, in natural water can be determined selectively by measuring the fluorescence lifetimes of the individual molecules. The different techniques for performing fluorescence lifetime measurements are described in Section II.C.3.

Remote fiber-optic Raman spectroscopy has found applications in process control. The principal advantage of this method is that many different molecules can be detected directly in solids, liquids, or gases, in complex media, or in harsh environments (i.e., high temperature or pressure). This technique is most sensitive to compounds that are IR-active. A large amount of qualitative and quantitative information can be gathered from the Raman spectrum without using any indicator chemistry. The low intensity of the Raman-scattered light is the major weakness of this technique. In addition, organic molecule fluorescence usually interferes with Raman spectra. This technique has been widely used for *in situ* monitoring of chemical reactions under harsh conditions. One application of Raman spectroscopy using optical fibers is to monitor epoxy-curing reactions. The extent of curing and the system temperature can be measured simultaneously. The degree of epoxy curing is calculated by taking the ratio of the epoxide ring stretch at 1240 cm^{-1}, which is linearly dependent on the progress of the curing reaction, and the 1186-cm^{-1} vibration of a fragment not affected by the cure. Raman spectroscopy with optical fibers is also used for on-line monitoring of water in sodium nitrate slurries in the nuclear industry. The Raman method is a nondestructive optical technique that can also provide detailed information about the molecular composition of tissues and it recently has been used for *in vivo* determination of the molecular composition of an arterial wall.

2. Gas-Phase Sensing

The most common fiber-optic gas sensor is based on spectral transmission analysis. Such analysis is commonly performed using two wavelength regions: 250–500 nm and ca. 1–8 μm. Absorption or emission in the lower wavelength region corresponds to electronic transitions within atoms or molecules, and is a useful region for measuring the energy changes associated with the transitions of a large number of gaseous species. The longer wavelengths cover the near- and mid-IR regions of the spectrum, and contain information about vibrational absorption bands of the gases. Conventional silica fibers are good for measuring the near-IR absorption lines but they are not good at the lower or mid-IR regions due to high attenuation and an increase in opacity of the silica, respectively. Therefore, a

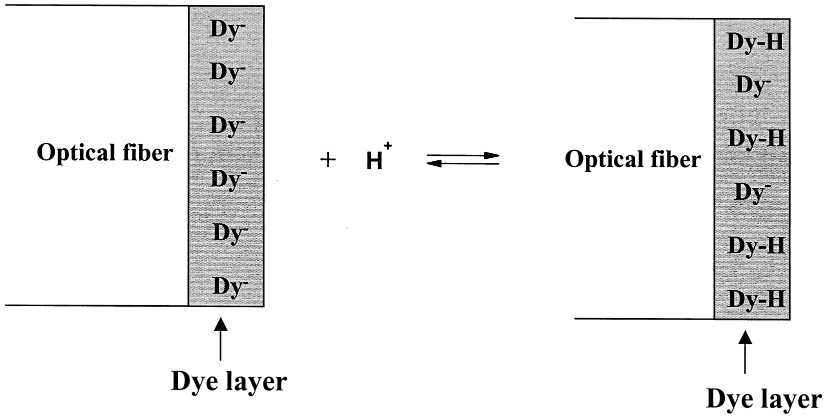


FIGURE 11 A fiber-optic pH sensor employing reversible protonation of dye (Dy^-).

long-path or multipass cell and a well-designed optoelectronic system are required reliably to detect low levels of gases. Although IR fibers can be used, they are fragile and expensive compared to conventional silica fibers. Direct absorption methods are mostly used to monitor low concentrations of gases such as nitrogen dioxide, methane, and hydrogen.

B. Extrinsic Sensing Mechanism—Indicator Chemistry

Extrinsic fiber-optic chemical sensors are constructed by immobilizing indicator chemistries on the fiber tip or on the annulus of the fiber. In this section, the mechanism of the chemical and biological reagents employed for sensing are described as well as methods for immobilizing sensing materials.

1. Chemical Sensing Reagent—Optrodes

a. pH sensing. Among all types of sensors, the pH sensor is the most widely developed and has received the most attention because of the importance of pH measurements in scientific research as well as various physiological, environmental, and industrial processes. Fiber-optic pH sensors offer several advantages compared to standard glass electrodes for pH measurement, including their immunity from electromagnetic interference, their small size, their ability to perform remote sensing, and their application to both *in vivo* and *in situ* measurements. Optical sensors also possess some disadvantages, such as photobleaching of the dye, leaching of dye from the immobilized surface, and interference by ambient light. The pH measurement is based on pH-dependent changes of the optical properties of an indicator dye immobilized on the fiber-optic surface. The indicator reversibly reacts with protons, and as a result, the absorption, reflection, or fluorescence properties of the indicator change. A schematic of a typical pH sensor is shown in Fig. 11. Some common dyes used for pH sensors are listed in Table I.

The fiber-optic, absorbance-based pH sensor monitors the absorbance change of the immobilized dye as a function of the pH change of the analytical medium. The Beer–Lambert law can be applied to an absorbance based pH sensor. The concentration of the dye [Dy] is related to the absorbance according to the following equation:

$$A = \log(I_0/I) = \varepsilon\,[\mathrm{Dy}]\,l, \qquad (5)$$

where I_0 and I are the intensity of transmitted light in the absence and presence of dye, respectively, l is the effective path length, and ε is the molar absorption coefficient. Usually the concentration of the base form of a weak acid indicator is measured, which is a measure of the pH-dependent degree of ionization of the indicator. The dissociation of a dye and the acidity constant K_a are expressed in the following equations, respectively:

$$\mathrm{Dy-H} = \mathrm{Dy}^- + \mathrm{H}^+, \qquad (6)$$

$$K_a = \frac{[\mathrm{Dy}^-][\mathrm{H}^+]}{[\mathrm{Dy-H}]}. \qquad (7)$$

Combining the equilibrium expression for the acid dissociation of a dye with the expressions for pH and pK_a gives

$$\frac{\text{Measured absorbance }(A)}{\text{Absorbance of the total dye in base form}} = \frac{1}{10^{(\mathrm{pK_a}-\mathrm{pH})}+1} \qquad (8)$$

This relationship results in a sigmoidal plot of absorbance versus pH centered on the $\mathrm{pK_a}$ value. In comparison to fluorescence, the absorbance method is simple

TABLE I Common Indicators

Indicator[a]	pK_a
Absorbance-based indicators	
Bromothymol blue	6.8
Chlorophenol red	6.3
Dibromo-xylenol blue	7.6
Neutral red	7.4, 5.9
Nitrazine yellow	6.5
Palatine chrome black	7.4
Phenol red	7.6
Phenoltetrachloro-sulfonaphthalein	7.0
Long-wave-absorbing pH indicators	
Methyl violet	0.0–1.6
Malachite green	0.2–1.8
Cresol red	1.0–2.0
Bromophenol blue	2.8–4.8
Naphtholbenzein	8.2–10.0
Alizarin yellow R	10.0–12.0
Alizarin	11.0–12.4
Indigocarmine	11.4–13.0
Fluorescence indicators	
Fluorescein	2.2, 4.4, 6.7
Eosin	3.25, 3.80
2′,7′-dichlorofluorescein	0.5, 3.5, 5.0
5(6)-carboxy-fluorescein	6.4
Carboxy naphthofluorescein	7.0
SNARF	7.6
SNAFL	7.6, 7.3

[a] SNARF, seminaphthorhodafluor; SNAFL, seminaphthofluorescein.

and easy to use, but it is not very sensitive, requiring the use of a high concentration of pH indicator and a relatively thick sensing layer.

Fluorescence-based fiber-optic pH sensors are more widely used due to their higher sensitivity. In this technique, the fluorescence intensity change of an immobilized dye is measured corresponding to a change in the medium pH. For example, the acid form of fluorescein does not fluoresce, but its conjugate base strongly fluoresces upon excitation. The concentration of deprotonated fluorescein is directly proportional to the measured fluorescence intensity and is dependent on the solution pH through its acid dissociation equilibrium. The intensity I_F of fluorescence light returning from the sensor tip is proportional to the concentration of the dye in the sensor and the intensity of the exciting radiation I_0,

$$I_F = k' I_0 \phi_F \varepsilon l [\mathrm{Dy}], \tag{9}$$

where l is the optical path length in the sensing layer, ε is the molar absorptivity, ϕ_F is the quantum yield of fluorescence, and k' is a constant related to the configuration of the instrument and sensor. When I_0 is constant, Eq. (9) can be simplified to

$$I_F = k[\mathrm{Dy}], \tag{10}$$

where $k = k' I_0 \phi_F \varepsilon l$.

Again the concentration of the dye can be easily related to the pH of the solution. The measured fluorescence intensity can be represented by the same form of equation as shown in Eq. (8),

$$\frac{\text{Measured fluorescence}}{\text{Fluorescence of the total dye in base form}} = \frac{1}{10^{(\mathrm{pK_a}-\mathrm{pH})} + 1}. \tag{11}$$

This relationship also results in a sigmoidal plot of intensity versus pH with a midpoint of the linear part of the curve corresponding to the pK_a of the immobilized dye. A pH sensor prepared by immobilizing a particular dye is useful over approximately two pH units (± 1 pK_a). Such a small pH range is a limitation of optical pH sensors compared to pH electrodes. A pH fiber-optic chemical sensor having a wide range of pH sensing capabilities can be constructed by immobilizing different dyes on a single fiber tip or a bundle of single-core fibers each containing a single dye with a particular pK_a value. When several dyes each having a different pK_a and a different optical spectrum are used, the pH in the region of each pK_a is determined separately by measuring the change in the distinctive spectrum for that pH region.

A fiber-optic pH sensor based on fluorescence energy transfer can be constructed by coimmobilizing a pH-sensitive fluorophore and a pH-sensitive absorber. For example, eosin (donor) and phenol red (acceptor) were coimmobilized in a polymer on the distal end of a silanized single-core optical fiber. Eosin's emission spectrum overlaps with the absorption of the basic form of phenol red. The concentration of the basic form of phenol red increases with an increase in pH. As a result, energy transfer from eosin to phenol red increases and the fluorescence intensity of eosin decreases. Thus, the pH-dependent absorption change of phenol red can be detected as changes in the fluorescence signal of eosin.

An evanescent-field-type pH sensor can be fabricated by replacing the cladding layer with a thin layer of pH-sensitive dye embedded in a polymer matrix. The basic designs are shown schematically in Fig. 8c. The measurement is based on the interaction of the evanescent wave with the dye in the coated cladding. A portion of the resulting dye fluorescence is coupled back into the fiber through the same mechanism that generates the original evanescent wave.

Other than dye-based indicators, certain conducting polymers can also be used for pH sensing. Conducting

polymers, e.g., polyaniline, substituted polyaniline, and polythiophene, vary in color as a function of pH. A thin coating of a conducting polymer on the end of the fiber surface can be used to detect the pH of a solution. A fiber-optic pH sensor has been developed based on absorbance changes of the polymers both in the visible and near-IR regions. These pH sensors have the advantage that the polymers can be used over a wide pH range since they are essentially electrolytes with multiple pK_a values. These polymers have certain disadvantages such as interference from other ions and the need for reconditioning (with HCl) before each measurement, which is necessary to nullify the conformational changes occurring in the polymer on pH changes.

Fiber-optic chemical sensors for different analytes can be developed based on pH sensors. Gases such as CO_2 and NH_3 react with water and change the pH of a solution when solvated, which can be detected by a pH sensor. Gas fiber-optic chemical sensors will be discussed in the next section. Any chemical or biological species that produces either acids or bases during a chemical or enzymatic reaction can be detected by measuring the pH of the medium.

b. Gas and vapor sensing. Gas sensing fiber-optic chemical sensors are mainly based on fluorescence quenching and acid–base chemistry. Optical oxygen sensors are primarily designed based on quenching of a luminescent dye. Fiber-optic chemical sensors for oxygen are constructed by immobilizing oxygen-sensitive fluorophores, using entrapment or adsorption on the fiber surface since most of these dyes do not have an appropriate functional group suitable for covalent immobilization.

The process of dynamic quenching is fully reversible, i.e., the dye is not consumed by reaction with oxygen. Dynamic quenching of the fluorophore is responsible for the decrease in fluorescence, and the Stern–Volmer quenching model can be used to express the extent of oxygen quenching:

$$I_0/I = 1 + K_{SV}[O_2], \tag{12}$$

where I is the fluorescence intensity at a particular oxygen concentration, I_0 is the value in the absence of oxygen, K_{SV} is the Stern–Volmer quenching constant, and $[O_2]$ is the concentration of oxygen. A linear calibration is thereby obtained by plotting the intensity ratio as a function of the oxygen concentration. At higher oxygen concentrations, this plot deviates from linearity.

Any dye whose fluorescence intensity is quenched by oxygen can be used to construct a fiber-optic oxygen sensor. Commonly used dyes include polyaromatic hydrocarbons (PAH) such as pyrene, fluoranthene, or benzoperylene, and organometallic complexes of ruthenium, osmium, palladium, and platinum. Typical polymeric supports used for immobilizing these indicators are silicone rubber, plasticized PVC, polystyrene, poly(hydroxyethyl methacrylate) (pHEMA), or porous glass. The sensitivity of the sensor depends on the dye interaction with oxygen and the gas permeability of the polymer film. Sensor selectivity is provided both by the molecular specificity of the quenching phenomenon and the molecular restrictions due to selective permeability of the polymer membrane.

Oxygen affects not only the intensity of the indicator fluorescence, but also its decay time. Organometallic dyes have longer fluorescence decay times and can be used to prepare oxygen sensors based on the measurement of this decay time. In comparison to fluorescence-intensity-based sensor types, decay-time-based sensors have less signal drift due to leaching or bleaching because the decay time is not dependent on fluorophore concentration.

Fiber-optic chemical sensors for acidic or basic gases such as ammonia, carbon dioxide, hydrogen cyanide, and nitrogen oxide can be constructed by coupling simple acid–base chemistry with pH sensors. First, pH-sensitive dyes or indicators are immobilized using polymers or sol–gel glasses and subsequently covered by a gas-permeable membrane on the distal end of a fiber. The gas-permeable membrane separates the sample solution from the immobilized dye. The acidic or basic gases cross the membrane, enter into the indicator layer, and undergo proton transfer with the dye. The extent of this reaction is monitored spectroscopically through the fiber.

For example, a fiber-optic *ammonia sensor* is prepared based on

$$\mathrm{Dy-H + NH_3 \rightarrow Dy^- + NH_4^+}. \tag{13}$$

Usually the nonprotonated form of the indicator dye is detected either by an absorbance or fluorescence measurement. As the ammonia concentration in the sample increases, the concentration of the nonprotonated dye increases, which causes an increase in the measured fluorescence or absorbance. The measured absorbance A is related to the ammonia concentration in the sample solution by the following expression:

$$A = \frac{\varepsilon b K_{\mathrm{eq}} C_{\mathrm{Dy}}[\mathrm{NH_3}]_{\mathrm{s}}}{C_{\mathrm{NH_3}} + K_{\mathrm{eq}}[\mathrm{NH_3}]_{\mathrm{s}}}, \tag{14}$$

where ε is the molar absorptivity of the chromophore, b is the effective path length at the sensor end, K_{eq} is the equilibrium constant of the above reaction, C_{Dy} and C_{NH_3} correspond to total dye and total ammonia concentration ($[NH_4^+] + [NH_3]$) in the dye solution, respectively, and $[NH_3]_s$ is the ammonia concentration in the sample solution.

Two different methods can be used to prepare a fiber-optic *carbon dioxide sensor*. In the first method, the

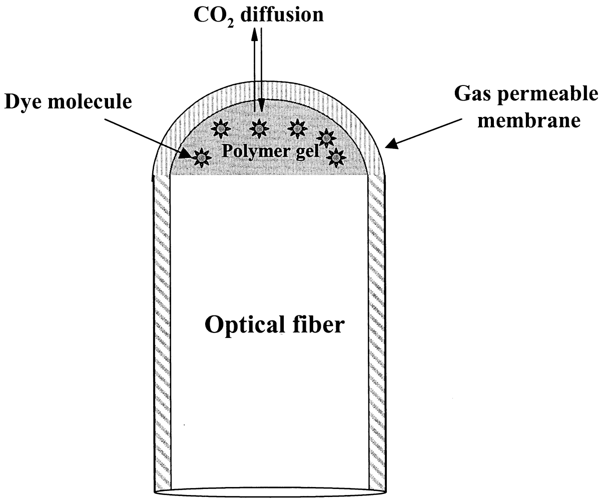


FIGURE 12 Fiber-optic sensor for carbon dioxide with a buffer solution entrapped in a polymer gel at the fiber tip and covered with a CO_2-permeable membrane.

sensors are obtained by entrapping a bicarbonate buffer along with an absorbance or fluorescence indicator into a polymer gel matrix and subsequently covering the matrix by a CO_2-permeable membrane at the fiber tip. A schematic arrangement of such CO_2 sensors is shown in Fig. 12. The sensor measures the pH of the bicarbonate buffer solution, which is in equilibrium with CO_2 outside the membrane, i.e.,

$$\begin{aligned} CO_2 &\overset{\text{membrane}}{\longleftrightarrow} CO_2 + H_2O, \\ CO_2 + H_2O &\Longleftrightarrow H_2CO_3, \\ H_2CO_3 &\Longleftrightarrow H^+ + HCO_3^-, \\ HCO_3^- &\Longleftrightarrow H^+ + CO_3^{2-}. \end{aligned} \qquad (15)$$

These reactions are governed by the following equilibrium constants:

$$K_1 = [H^+][HCO_3^-]/[CO_2], \qquad (16)$$

$$K_2 = [H^+]\left[CO_3^{2-}\right]/[HCO_3^-]. \qquad (17)$$

By considering the above two equilibria, the equation used to relate the external $[CO_2]$ to the internal $[H^+]$ is

$$[H^+]^3 + N[H^+]^2 - (K_1[CO_2] + K_W)[H^+] - 2K_1K_2[CO_2] = 0, \qquad (18)$$

where N is the bicarbonate ion concentration in the internal solution, $[CO_2]$ is the total analytical concentration of carbon dioxide, i.e., $[CO_2] = [CO_2]_{(aq)} + [H_2CO_3]$, and $K_W = [H^+][OH^-]$. As carbon dioxide crosses the membrane, the microenvironmental pH of the entrapped buffer changes, changing the absorbance or fluorescence intensity of the immobilized dye. The concentration of CO_2 is determined by measuring the pH of the buffer solution.

The second type of CO_2 fiber-optic chemical sensor is constructed by using ion pairs consisting of a pH indicator anion and an organic quaternary cation. First, a pH indicator dye (DH) and a quaternary ammonium hydroxide (Q^+OH^-) are entrapped into a proton-impermeable but CO_2-permeable polymer membrane, which is then immobilized onto the fiber's surface. The mechanism of this CO_2 sensor is based on the interaction between the dye molecules (DH) and the quaternary cations (Q^+OH^-) to form hydrated ion pairs ($Q^+D^- \cdot xH_2O$). The hydrated ion pair is dissolved in the polymer, where it reacts with CO_2 according to the following reaction:

$$\{Q^+D^- \cdot xH_2O\} + CO_2 \Longleftrightarrow \{Q^+HCO_3^- \cdot (x-1)H_2O\} + DH. \qquad (19)$$

The indicator dye becomes protonated and changes its absorption (or emission) maximum. Such sensors can be used for determining CO_2 in dry gases as well as in aqueous solutions.

Fiber-optic chemical sensors for the detecting *organic vapors* (such as methanol, chloroform, benzene, toluene, and nitroaromatics) are constructed by immobilizing a solvatochromic fluorescent dye into various polymers on the fiber tip. Solvatochromic indicators change their fluorescent properties (such as intensity, emission wavelength maximum) as the polarity of the medium changes. For example, the solvatochromic indicator Nile red exhibits a large shift in emission wavelength maximum with changes in local polarity. Typically, its absorption and/or emission spectra shift to higher wavelengths when exposed to solvents with increasing polarity. A single solvatochromic dye can be immobilized within a polymer matrix, having its own baseline polarity. The sorption of organic vapors into the polymer changes the microenvironmental polarity of the dye, which changes its emission/absorption maximum. Different organic vapors change the polarity, hydrophobicity, and swelling tendency of the polymer, generating different fluorescence responses of the immobilized dye.

c. Ion sensing. Several different schemes can be applied to fiber-optic chemical sensors for detecting ions other than hydrogen. One approach is to design a system, similar to pH fiber-optic chemical sensors, in which a dye that selectively binds a metal ion of interest is immobilized in an ion-permeable polymer such as cellulose or a hydrogel at the tip of an optical fiber. The reaction between the dye and the ion changes the absorbance or fluorescence of such dyes. Absorbance or fluorescence intensity changes are measured as a function of ion concentration, but this

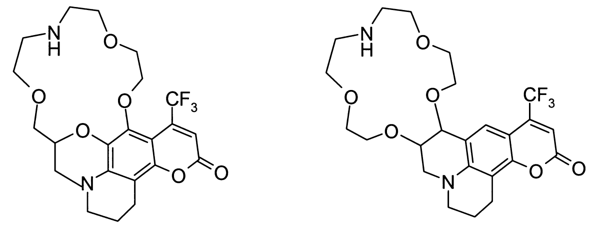


FIGURE 13 The chemical structure of two dyes that possess metal-binding crown ethers. The effects of metal ion chelation will have different photophysical consequences depending on the location of the interaction. This interaction will cause the dyes to display unique emission spectra upon metal binding.

approach is not as selective as other methods (discussed below).

The second ion-sensing approach involves the use of fluorogenic and chromogenic crown ethers attached to the fiber tip directly or via an ion-permeable membrane. Fluorogenic and chromogenic crown ethers are prepared by covalently attaching a fluorogenic or a chromogenic dye to crown ethers (some examples are given in Fig. 13). The molecular design of these crown ethers must assure that the spectroscopic properties of the attached dye change when the crown ethers bind to metal ions. Ion-sensitive fiber-optic chemical sensors prepared with crown ethers are highly selective due to chemical recognition for specific metal ions. The sensitivity of this type of fiber-optic chemical sensor for detecting ions in an aqueous environment is relatively low since the formation constants for metal ions binding with the crown ether in water are much lower than in nonaqueous environments.

There are two other commonly used schemes for preparing ion-sensitive fiber-optic chemical sensors. The first scheme is based on an ion-exchange technique. A complexing reagent (a cation-selective neutral ionophore) along with a spectroscopically detectable coreagent (fluorescent or absorbant anionic dye) is immobilized inside a thin hydrophobic membrane attached to the fiber. The operating mechanism of this sensor is based on electroneutrality. When analyte ions enter the membrane and selectively bind with the ionophore, an equal number of protons must be released from the membrane (see Fig. 14a). The indicator dye within the membrane acts as the proton donor, thereby altering the measured absorbance or fluorescence. The optical properties of the dye are thereby modulated by the extent of binding of the analyte cation. A constant pH level is maintained by using appropriate buffers.

An alternative ion-sensing scheme is referred to as coextraction. In this technique, a highly lipophilic anion such as chloride or salicylate is extracted into the membrane along with a cation, which is usually a proton to maintain electroneutrality. The highly lipophilic nature of the anions makes them suitable for extraction by the lipid membrane. Indicator dyes present inside the membrane become protonated and, as a result, the optical properties of the dye change. These changes are easily correlated to the anion concentrations in the aqueous phase. A schematic representation of this sensing scheme is shown in Fig. 14b. Both of these approaches (ion-exchange and coextraction), however, are strongly dependent on pH, which makes them hard to apply to samples where the pH is undefined.

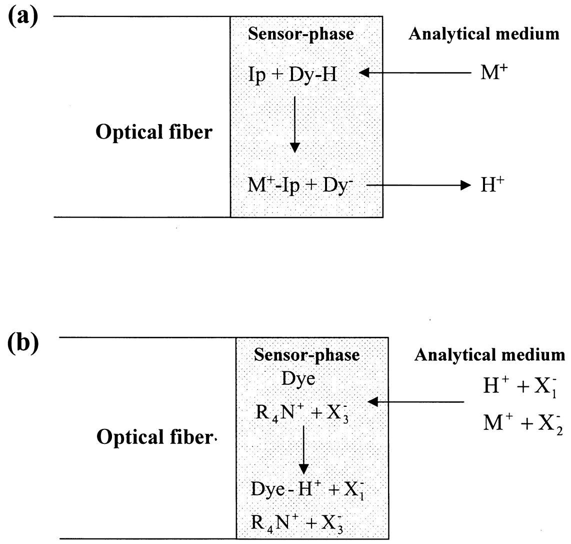


FIGURE 14 (a) Schematic of the mechanism of proton exchange for a metal ion inside a thin polymer membrane containing an ionophore and a protonated dye. (b) Schematic of the coextraction procedure. X_1^- is more lipophilic than X_2^-, hence X_1^- is more extractable.

2. Immobilization Techniques

Immobilization of sensing materials such as indicators or dyes is a key step in fiber-optic chemical sensor development. The sensing materials employed will largely determine the sensor characteristics for a particular application. Reagent immobilization procedures may involve several steps. The number of immobilization steps should be minimized to maximize yield. A good immobilization method should be (a) simple, (b) fast, (c) general, i.e., the immobilization method can be employed for a variety of sensing materials, (d) stable, i.e., the reagents do not leach from the substrate, and (e) gentle so as to retain the chemical and/or biochemical activities of the species being immobilized. There are three main methods for immobilizing an indicator: adsorption/electrostatic, entrapment, and covalent binding. A schematic representation of these methods is shown in Fig. 15.

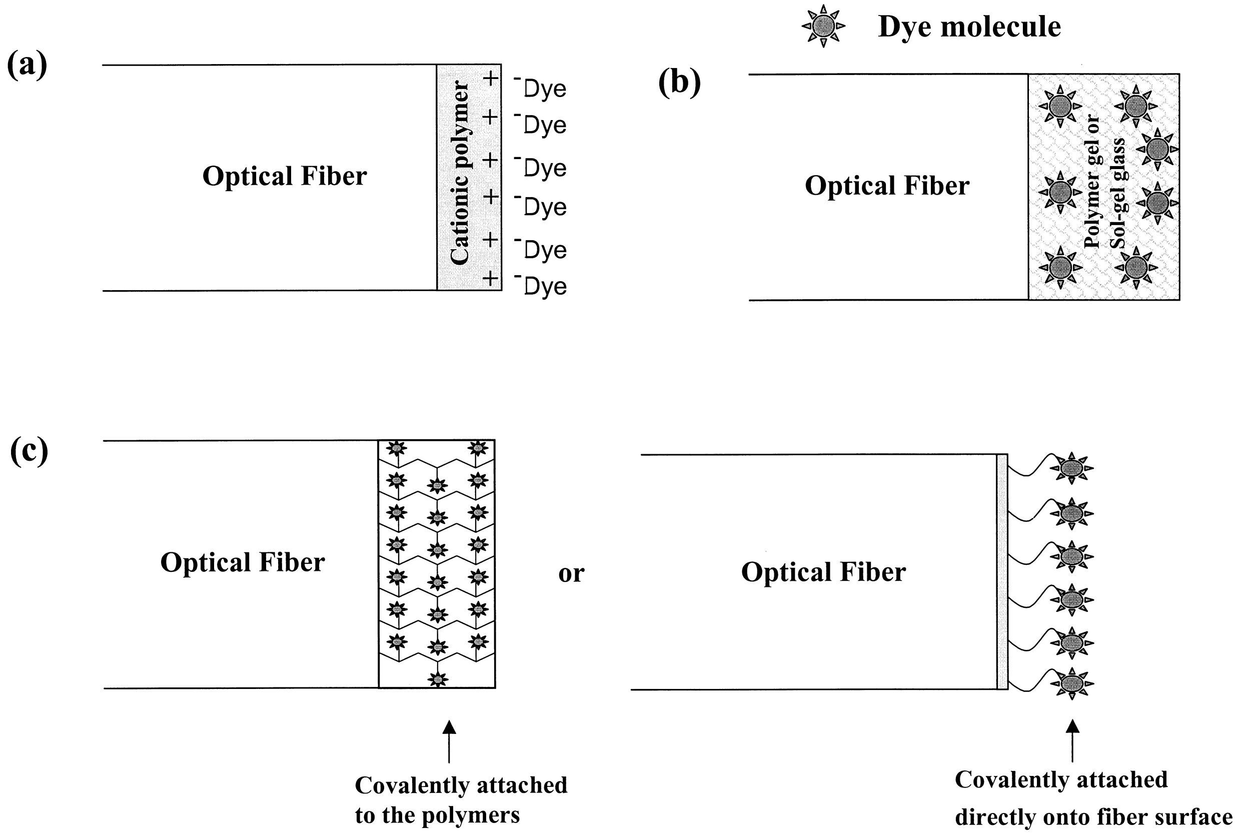


FIGURE 15 Schematic diagram of three different immobilization techniques. (a) Absorption/electrostatic, (b) entrapment, (c) covalent immobilization.

a. Adsorption/electrostatic immobilization. Adsorption immobilization methods involve adsorbing the sensing material onto a solid surface or polymer matrix. Sensing materials can be adsorbed onto the end of optical fibers either directly or with the aid of a thin polymer film. This immobilization method is the simplest of all the methods (described below); however, the adsorbed sensing material tends gradually to leach out from the solid support, decreasing sensing performance and/or lifetime. The surface may first be modified with complementary functional groups to enhance retention of the material. For example, a hydrophobic surface can be prepared to immobilize a hydrophobic species.

Electrostatic immobilization employs similar complementarity. The fiber surface can be coated with a thin polyelectrolyte membrane, which interacts electrostatically with oppositely charged sensing materials. For example, a negative surface, such as sulfonated polystyrene, can be prepared to efficiently immobilize a positively charged dye. Similarly, cationic polymers can bind various anions. Ionic indicators or biological molecules (i.e., either cations or anions) can be immobilized by electrostatic interaction. In order to prevent leaching, strong ion exchangers are employed.

The absorption immobilization method is very easy and highly reproducible. Sensing material loading can be controlled easily by changing the immobilization time. Also, the immobilized reagents are easily accessible to analyte because they are situated at sites on the polymer surface.

b. Entrapment immobilization. In the entrapment method, the sensing material is physically trapped within a porous matrix. Many different indicators, such as fluorophores or biological materials, may be immobilized in this fashion. When this immobilization method is designed properly, trapped reagents either do not leach out or leach out very slowly from the matrix. The rate of leaching depends on the indicator size, molecular weight, and porosity of the matrix. Organic polymer matrices have been widely used for entrapping sensing materials in the void volume of polymers. Sensing materials are first dissolved in a common solvent with an appropriate polymer or polymer precursor and then the optical fiber is either spin-coated or dip-coated with the viscous solution. Upon

curing or cross-linking, the polymer forms and entraps the material within its pores. Alternatively, complementary solubility between polymer and sensing material can be used to effect the entrapment. For example, lipophilic indicators will dissolve in lipophilic polymers and are not easily leached. Another way to entrap an indicator is to initiate polymerization of a monomer solution containing the indicator. When the polymer is formed, it entraps the indicator. Such polymers can be either thermally or photochemically initiated and attached to the fiber surface by dip-coating procedures. Silanization of the fiber surface with polymerizable groups, such as acrylates or vinyl residues, enhances adhesion of the polymer to the surface.

Optically transparent sol–gel glasses are also used for sensing material entrapment. Sol–gel glasses are produced by hydrolysis and polycondensation of organometallic compounds, such as tetraethyl orthosilicate, $Si(OCH_3)_4$. A sensing material is added to the reaction mixture at some time during the formation of the sol or gel. This viscous sol–gel solution is coated onto an optical fiber to form a sensing element. Sol–gel glasses prepared by this method contain interconnected pores formed by a three-dimensional SiO_2 network. As a result, the sensing material is trapped but small analytes can readily diffuse in and out of the pores. One advantage of sol–gel glass immobilization is its compatibility with many inorganic and organic reagents allowing many types of sensing materials to be entrapped. Also, sol–gel glasses are chemically, photochemically, and mechanically stable and solvent resistant, and can therefore be useful in harsh conditions. Disadvantages of sol–gel glass immobilization are the slow response times in aqueous media and the fragility of the thin sol–gel glass films compared with polymer films.

c. Covalent immobilization. Robust sensing materials are formed by covalent binding to the substrate. Both the substrate and the sensing materials must contain reactive groups for covalent attachment to occur. There are many methods for covalently immobilizing a sensing material to a fiber surface. One simple way is to covalently modify the fiber surface by silanizing with trialkoxysilyl compounds of the type $(RO)_3SiR'$, with R being ethyl or methyl and R′ being aminopropyl, 3-chloropropyl, 3-glycidyloxy, vinyl, or a long-chain amine. The functional group on the fiber surface then reacts with the sensing materials. In some cases, the sensing material must first be activated for reaction with the substrate. For example, dyes possessing —COOH groups can be converted to *N*-hydroxysuccinimidoyl esters that can react with an amine-modified surface. Amine-modified surfaces can also be derivatized with amine-containing dyes by using bifunctional cross-linkers such as glutaraldehyde or by direct reaction with an isothiocyanate-containing dye. Sensing materials are also covalently immobilized to a fiber surface by using similar chemistries to those employed with thin polymer or sol–gel glass films.

An alternative method for dye immobilization is based on photopolymerization. Dye-doped monomers or dyes containing polymerizable groups can be copolymerized with a monomer on a silanized fiber tip. In these procedures, it is important not to modify the sensing material in such a way as to disturb its ability to bind to the analyte or transduce the signal. Covalent immobilization methods are usually complicated and time-consuming compared with the other two immobilization techniques, but are very reliable since the dye or indicator is not likely to leach out. Covalent binding may result in a reduced response of the sensing material due to bond formation and reduced degrees of freedom upon immobilization.

3. Biological Sensing Materials (Fiber-Optic Based Biosensors)

Fiber-optic biosensors are a subtype of fiber-optic chemical sensors that rely on sensing materials of biological origin. In fiber-optic biosensors, biological recognition components such as enzymes, antibodies, DNA, or cells are attached to the optical fiber sensing layer in order to alter the specificity of the sensor.

In nature, interaction between biological molecules, such as receptor–ligand, antibody–antigen, or two complementary DNA strands, are highly specific. Some of these recognition molecules can be purified and used in fiber-optic biosensors. Moreover, by using genetic engineering, recombinant recognition molecules can be produced and used. Fiber-optic biosensors can be miniaturized and used in portable analytical devices for clinical, environmental, and industrial applications. Clinical fiber-optic biosensors have been developed for detecting cancer cells, pathogenic bacteria, viruses, toxic proteins, hormones, and drugs. Environmentally important analytes such as pesticides, heavy metals, and carcinogenic compounds can also be detected using fiber-optic biosensors. Industrial applications of fiber-optic biosensors include on-line process monitoring of bacteria or mammalian cell-based bioprocesses.

Biological sensing molecules enable the detection of an expanded number of analytes. Such molecules tend to be sensitive to pH or temperature changes and have poor stability, resulting in short lifetimes. Another important limitation is the high cost of purified biological sensing compounds. Biological sensing compounds can be divided into two categories based on their bioactivity: biocatalysts (enzymes and cells) and bioaffinity molecules (antibodies, receptors, and nucleic acids).

a. Biocatalysts as the recognition element in fiber-optic biosensor. Enzymes are the most commonly employed biocatalysts for fiber-optic biosensor fabrication. These proteins selectively catalyze the conversion of a substrate to product. In fiber-optic biosensors, enzymes are used to catalyze the conversion of a non-optically detectable analyte (the analyte is the enzyme's substrate) into an optically detectable product. The optical signal obtained, for example, absorbance or fluorescence, is proportional to the product concentration and thereby to the analyte's concentration. The enzymatic reaction products are measured either directly, if they are optically detectable, or indirectly by using indicators as described in Section III.B.1. Indicators are used to measure common reaction products such as H^+, ammonia, oxygen, carbon dioxide, and hydrogen peroxide. An example of this approach is the glucose biosensor. The enzyme glucose oxidase catalyzes the oxidation of glucose with oxygen (the substrates) to produce gluconolactone and H_2O_2. The glucose concentration is determined by using an indicator to measure either the amount of oxygen consumed or the amount of H_2O_2 produced using an appropriate indicator,

$$\text{Glucose} + O_2 \xrightarrow{\text{glucose oxidase}} \text{Gluconic acid} + H_2O_2.$$

The use of enzymes as sensing materials for fiber-optic biosensors is relatively limited since the stability and the activity of most enzymes significantly decrease when they are removed from their natural environment inside the cells, although immobilization techniques may help. To overcome this stability problem, whole cells may be used as biocatalysts. Although some of the fiber-optic biosensor specificity is reduced with whole-cells, the enzymes within the cells function more efficiently because they are in an optimum environment containing all the necessary cofactors. Whole-cell biocatalysts are particularly advantageous when the detection is based on a sequence of multiple enzymatic reactions. These enzymatic cascade reactions are very difficult and complicated to accomplish by coimmobilizing the enzymes but are relatively straightforward by employing whole cells. The methods for detecting the products in cell-based fiber-optic biosensors are similar to those employed in enzyme fiber-optic biosensors.

b. Bioaffinity as the recognition mechanism in fiber-optic biosensors. Antibodies, receptors, and nucleic acids are highly selective and can recognize their binding partners. When the affinity binding reaction is in equilibrium and readily reversible, a sensor results. If the reaction is essentially irreversible, the resulting sensor is called a probe, as it must be regenerated or recharged before it can be used to make another measurement. Most bioaffinity fiber-optic biosensors are based on transducing antibody–antigen (analyte) interactions into an optical signal that is proportional to the antigen concentration. Several detection schemes are employed; the simplest one involves the detection of intrinsically fluorescent analytes such as polynuclear aromatic hydrocarbons (PAHs). Antibodies are immobilized on the fiber surface and a fluorescence signal is obtained when the analyte (antigen) is present, as shown in Fig. 16a. A more generalized detection scheme, often called a competition assay, is based on competition for the antibody-binding site between an externally added fluorescent-labeled antigen and the antigen present in the sample (analyte), as shown in Fig. 16b. The antibody is immobilized on the optical fiber surface, a known concentration of fluorescent-labeled antigen is added, and the fluorescence signal obtained is set as the initial signal. To perform an analysis, the same fluorescent-labeled antigen concentration is mixed with a sample containing an unknown antigen concentration.

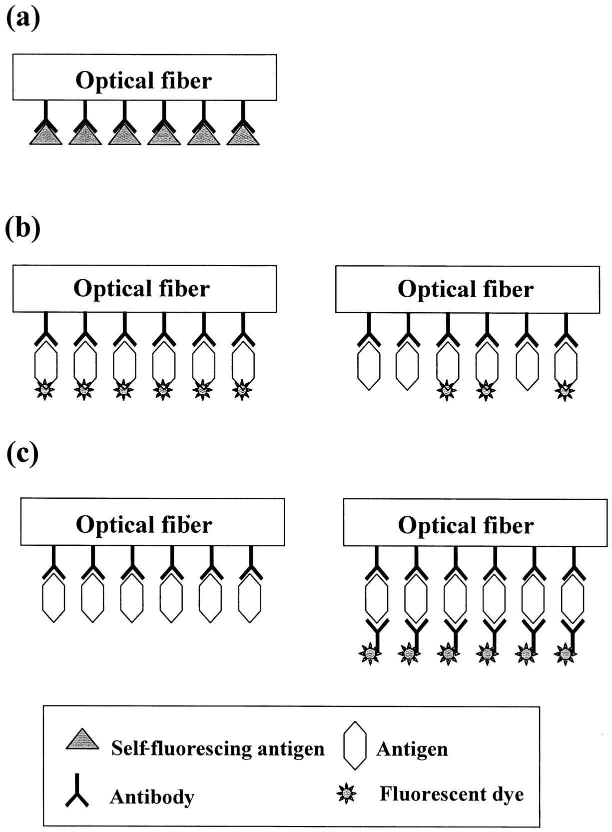


FIGURE 16 Schematic principles of bioaffinity fiber-optic biosensors. (a) Detection of intrinsically fluorescent molecule using immobilized antibody. (b) Competition assay using a fluorescent-labeled antigen. (c) Sandwich immunoassay using an immobilized antibody and a fluorescent-labeled antibody.

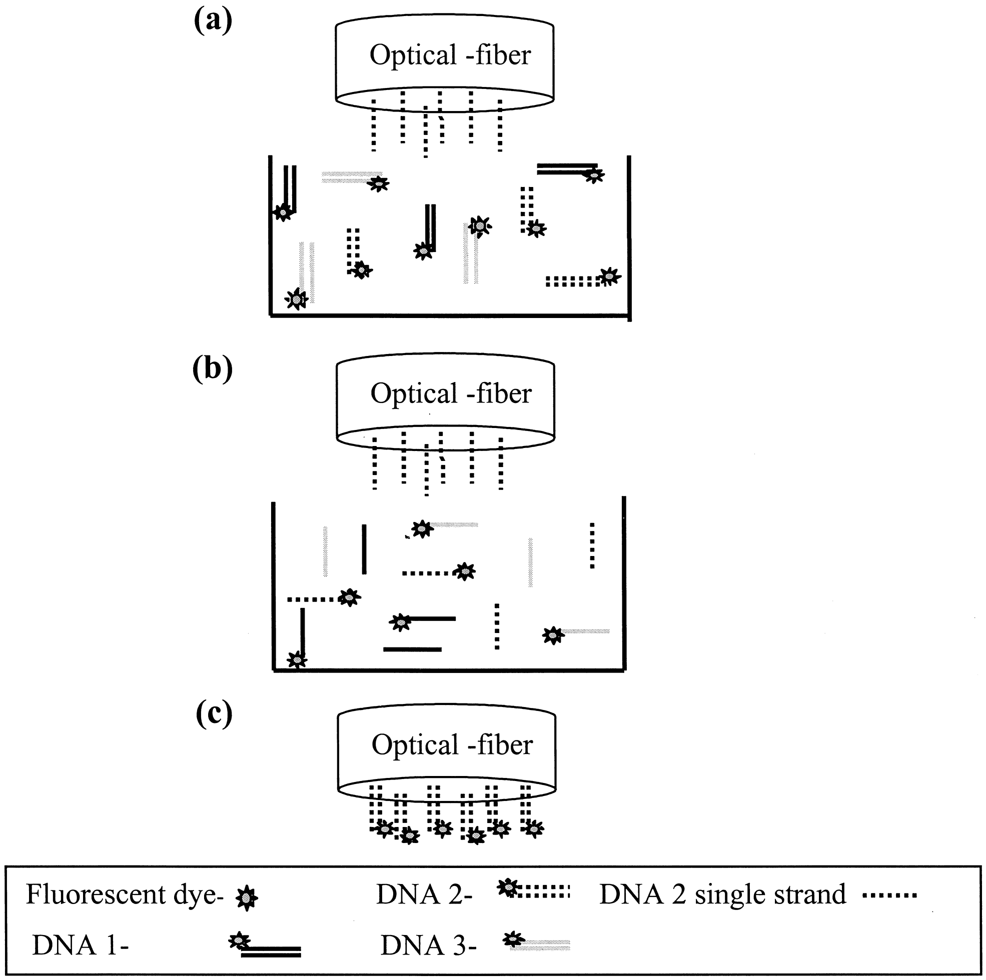


FIGURE 17 Principle of DNA fiber-optic biosensors. (a) Single-strand DNA probe molecules, with a sequence complementary to one strand of the target DNA sequence, are immobilized onto the fiber. (b) The fluorescent-labeled sample DNA molecules are first dehybridized and the fiber is dipped into the sample solution. (c) After hybridization, the complementary strands of the target DNA are attached to the probe DNA on the fiber and a fluorescence signal is obtained.

When this mixture is analyzed using the fiber-optic biosensor, the resulting fluorescence signal obtained is lower than the initial signal. The relative decrease in the initial signal is proportional to the analyte concentration in the sample.

A sandwich immunoassay is another widely used detection scheme and involves the use of two antibodies. The first antibody is immobilized on the fiber and is used to capture the antigen, and the second antibody, which is conjugated to a fluorescent dye or enzyme, is used to generate the signal (Fig. 16c). When an enzyme is used for antibody labeling, the enzymatic conversion of a nonfluorescent substrate to a fluorescent product is measured. The enzyme-labeling method is more sensitive since the signal is amplified by the enzymatic reaction.

Nucleic acid base pairing can also be used for bioaffinity sensor fabrication. The presence of a specific DNA sequence, the "target," among millions of other different sequences is detected by hybridization to its complementary DNA sequence, the "probe," which is immobilized on the optical fiber, as shown in Fig. 17. The sample DNA is labeled using fluorescent primers and the polymerase chain reaction (PCR). The resulting double-strand DNA molecules are dehybridized (usually by heating) and then allowed to rehybridize (by cooling) to the single-strand DNA probe molecules immobilized on the fiber surface. If the complementary target DNA sequence is present in the sample, a fluorescence signal is detected on the sensor. The target sequence can be, for example, a unique sequence found only in specific pathogenic bacteria. DNA

can be easily extracted from water, wastewater, or clinical samples and the presence of pathogenic microorganisms can be determined by the sensor. In general, such DNA-sensing schemes are rarely conducted using a single probe sequence but are usually performed with an array of hundreds to thousands of probes.

IV. APPLICATIONS OF FIBER-OPTIC CHEMICAL SENSORS

A. Clinical Applications

Clinical analytical devices are important for various diagnostics applications since the use of these devices can directly lead to more efficient and effective medical treatment. Most diagnostic tests are performed in a centralized laboratory. Samples must be collected with the attendant transport, storage, and chain-of-custody issues. The remote location of the laboratory delays the medical diagnosis. For these reasons, analytical devices such as sensors that can be located near the patient's bed or even inside the patient's body are of great utility. In recent years, fiber-optic chemical sensors have been developed for both *in vivo* diagnostics (monitoring physical and chemical parameters inside the body) and for clinical sample analysis at the patient bedside (i.e., blood or urine tests). The principal advantage of fiber-optic chemical sensors over electrochemical technologies is that no electrical current is involved in the measurement. Thus, the measurement inside the body is much safer, and analysis at the patient's bedside does not affect other electrical medical devices.

1. Fiber-Optic Chemical Sensors for *In Vivo* Analysis

In vivo analytical devices ideally should be capable of monitoring several different physiological parameters simultaneously without interfering with an ongoing medical procedure, such as surgery. The devices should be biocompatible, simple to implement and operate, and highly reliable and safe. Fiber-optic chemical sensors can meet most of these requirements since the optical fibers are small (few hundred micrometers in diameter), flexible, nontoxic, and chemically inert. Optical fibers have already proven to be valuable for *in vivo* clinical applications such as endoscopic procedures and laser power transmission for surgical applications.

Typical fiber-optic chemical sensors for *in vivo* monitoring are constructed from optical fibers that are connected to an external compact unit containing the light source and the detector, as shown in Fig. 18. The optical fiber's distal end is inserted into the patient through a catheter.

Fiber-optic chemical sensors were originally used to measure oxygen saturation in the blood. The oxygen carrier in the blood is hemoglobin (Hb) and the blood saturation is the ratio (as a percentage) between the oxygen that is actually carried by Hb and the Hb's maximum carrying capacity. The measurement is based on the large difference in the light absorption of Hb and OxyHb (oxyhemoglobin, i.e., oxygen bound to Hb) at $\lambda = 660$ nm. Typically the light is delivered through two bifurcated fibers at two different wavelengths, 660 nm (where Hb absorption is higher than that of OxyHb) and 805 nm (where there is a small difference in absorption between Hb and OxyHb); the absorption is measured and the blood saturation is calculated.

Fiber-optic chemical sensors have been developed for measuring all three blood gas parameters pH, PO_2, and PCO_2. These parameters are used to determine the efficacy of gas exchange within tissue and are crucial in surgical procedures such as heart bypass surgery and critical care monitoring for patients with compromised respiratory conditions. The sensing mechanism is based on dyes that change their fluorescence intensities as a function of analyte concentration (see Section III.B.1). Several sensor configurations have been developed for intravascular measurement including one that incorporates the three different sensors into a single device. At present these sensors have not been implemented because of blood compatibility problems in which a thrombus (clot) forms around the sensor tip and affects the measurement accuracy.

Several pH *in vivo* fiber-optic chemical sensors for other clinical applications have been developed. These include pH monitoring in the stomach and the detection and examination of malignant tumors. Both sensors are based on pH-sensitive fluorescent dyes. In gastric sensors, the dyes are immobilized on the fiber tip and the emitted light intensities are correlated to pH changes. Since a wide pH range (1–8) is monitored, there is a need to use several different dyes (each for a different pH range). pH sensors have also been used for locating suspicious tissue and determining if malignancy is present. This oncological sensor is based on the observation that malignant tumors can induce a decrease in their microenvironment pH. Thus, by inserting the optical fiber into the tissue and then injecting a nontoxic pH indicator (fluorescein based), it is possible to identify and map malignant tumor locations.

An interesting fiber-optic chemical sensor device for *in vivo* gastric diagnostics is the Bilitec 2000 (Medtronic Synectics AB). This device is used to measure the bilirubin concentration in the stomach and the esophagus. The bilirubin concentrations are related to bile-containing reflux in these organs and can reveal several pathological conditions such as gastric ulcers and gastric cancer. The device, shown in Fig. 18, consists of two light-emitting

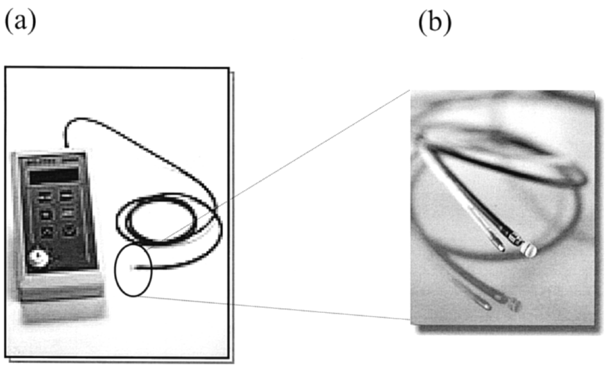


FIGURE 18 Bilitec 2000 fiber-optic chemical sensor for *in vivo* gastric diagnostics. (a) The external unit and the optical fibers. (b) The optical fiber's distal end that is inserted into the body. [Reproduced with permission from Medtronic Synectics AB.]

diodes that emit light at $\lambda = 465$ nm (bilrubin absorption) and 570 nm (reference). The light is transmitted by the fiber bundles to the probe (miniaturized spectrophotometric cell) at the fiber's distal end. Bilirubin's light absorption is measured by the detector and bilirubin concentrations are calculated.

Several clinical fiber-optic chemical sensors are commercially available and others are in different stages of development. The demand for *in vivo* fiber-optic chemical sensors for monitoring important analytes such as glucose, potassium, urea, lactate, and some enzymes has led to concerted research efforts in this field.

2. Clinical Fiber-Optic Chemical Sensors for *In Situ* Sample Analysis

Fiber-optic chemical sensors can be located at the patient's bedside or even in the patient's home for self-use. Such sensors are designed as compact, simple–to-use devices. While these sensors employ essentially the same sensing chemistry as with *in vivo* sensors, the overall sensor design can be much simpler and a wider range of materials may be used for sensor fabrication. Fiber-optic chemical sensors for blood gases (PO_2, PCO_2, and pH) are commercially available. Measurements are based on the use of immobilized fluorescent dyes. In one device, the dyes are incorporated into a disposable apparatus that is inserted into an extracorporeal blood circuit on one side and connected to a fiber bundle on the other. These sensors are mainly used to monitor blood gases during open heart surgery.

Another approach is to measure pH and blood gases using a paracorporeal measurement at the patient's bedside. The sensors are three fluorescent-sensing materials responsive to each of the blood gases. The sensors are placed into an external tube connected to an arterial blood line. Blood samples are periodically and automatically pumped into the tube, analyzed by the sensors, and then returned to the blood line. In this way, the blood can be monitored semicontinuously without requiring blood samples to be taken from the patient.

Fiber-optic biosensor can be potentially used for clinical sample analysis. As described above, by using enzymes and antibodies immobilized on the fiber tip or around the fiber core (evanescent field), different fiber-optic biosensors have been developed. In most cases, the measurement is based on changes in fluorescence intensity. The target analytes include glucose, cholesterol, enzymes, antibodies, bacteria, and viruses. Measurements are usually fast and simple, and in many devices, the probe is disposable; however, the instability of the biological recognition materials reduces the sensor lifetime.

Recently, fiber-optic biosensors for detecting DNA sequences have been reported. These fiber-optic biosensors can be used for detecting pathogenic microorganisms and for identifying defective genes. The sensors are usually designed in an array format (Fig. 19) allowing the simultaneous analysis of hundreds or thousands of samples. Although just recently developed, these sensors will undoubtedly revolutionize the clinical diagnostics field.

B. Environmental Fiber-Optic Chemical Sensors

Both short- and long-term harmful effects of toxic chemical accumulation in the environment have become apparent in the last few decades. In order to reduce the environmental damage caused by these pollutants, regulations were set that either restricted or completely prohibited the release of environmentally hazardous chemicals into the environment. The efficient enforcement of these regulations is highly dependent on sensitive and reliable analytical tools for environmental monitoring. These analytical tools ideally should be sensitive, simple to construct, portable, and suitable for continuous field (i.e., *in situ*) measurements. Fiber-optic chemical sensor systems can meet most of these requirements since they can be miniaturized and integrated into portable analytical devices that can offer high specificity and sensitivity, fast response, mechanical stability, and low cost. Furthermore, by using long optical fibers, the sensors can be used for remote analytical monitoring.

Most commercially available environmental fiber-optic chemical sensors are for pH and oxygen monitoring in water and wastewater. The monitoring of unusual changes in oxygen and pH can be used as an indirect indication of the presence of pollutants. These sensors are based on fluorescent dyes as the sensing material. The optical fiber and the sensing elements are covered with a metal jacket that provides immunity from harsh environmental conditions.

Environmental sensors capable of detecting specific pollutants such as volatile organic compounds and heavy metals (two of the largest classes of chemical pollutants)

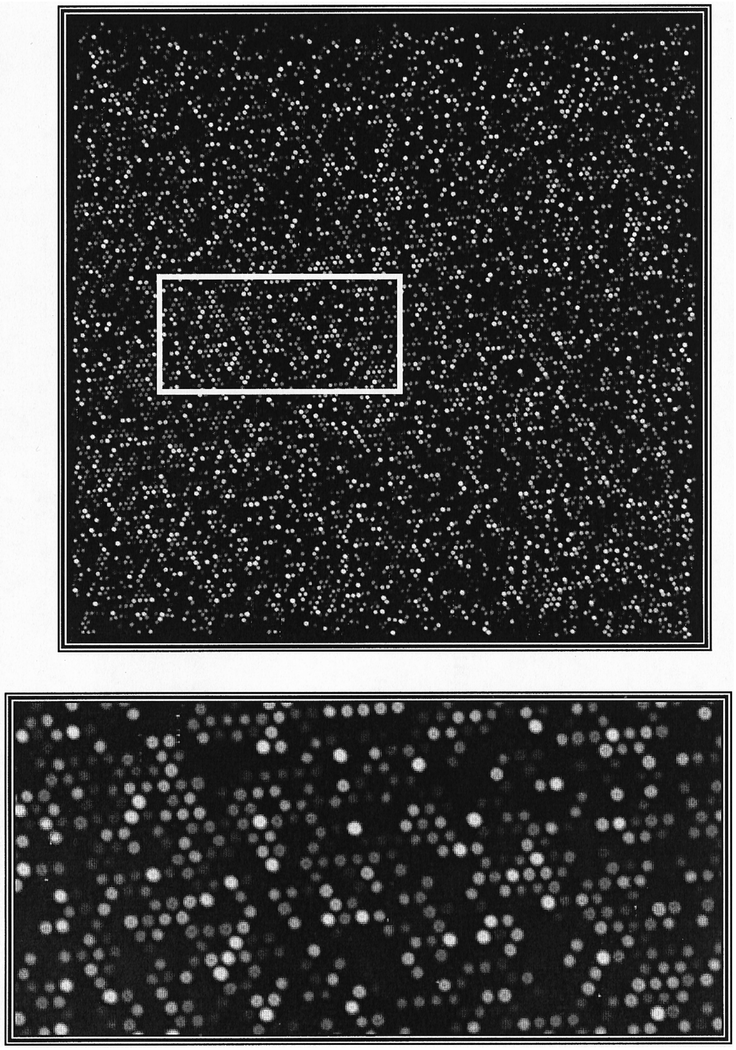


FIGURE 19 Fluorescence image of a DNA sensor array with ~13,000 DNA probe regions.

have also been developed. For detecting volatile organic compounds, direct fluorescence spectroscopy can be used since most volatile organic compounds have unique fluorescence spectra. Moreover, by using time-resolved fluorescence spectroscopy, it is possible to distinguish between volatile organics with similar structures, for example, toluene and benzene. Heavy-metal fiber-optic chemical sensors are mostly based on ion sensitive indicators as described in Section III.B.1.c.

Pesticides are another important group of pollutants that can be detected by fiber-optic chemical sensors. Since pesticides are designed to interact with biological molecules, fiber-optic biosensors are mostly used for their detection. One example is the detection of organophosphate and carbamate pesticides by monitoring their inhibition effect on the enzymatic reaction of acetylcholinesterase (AChE) with its substrate, acetylcholine. The enzyme is coimmobilized at the distal end of the fiber together with a pH-sensitive dye. The formation of acetic acid, which is one of the enzymatic reaction products, changes the local pH and thereby the fluorescence signal. The inhibition of this reaction can be related to the pesticide concentration in the sample.

Current research activities on environmental fiber-optic chemical sensors are focused on developing multianalyte sensors for on-line and *in situ* monitoring. These sensors may one day be used as environmental warning devices that will help prevent environmental catastrophes by triggering an alarm whenever uncontrolled release of pollutants to the environment occurs.

C. Industrial and Bioprocess Control Applications

In the chemical and bioprocess industries, the need for real-time process and quality control has elicited a great

interest in chemical sensors. Effective process control is a basic requirement for the optimal utilization of chemical and biological materials. Conventional analytical techniques for measuring industrial or biological materials such as GC, HPLC, and flow injection analysis have several drawbacks such as price, large instrument size, interference by medium components, and drift.

Intrinsic fiber-optic chemical sensors are widely used in industry and bioprocess control. Two basic approaches are used in remote applications. In the first approach, the spectroscopic parameters are chosen such that the analyte gives a unique signal compared with all other components in the sample. Examples of such approaches are as follows: (a) Fiber-optic remote fluorescence spectroscopy is used to measure reduced nicotinamide adenine dinucleotide (NADH) in bioreactors. The idea is to correlate the viability of a population of cells within the reactor to the total amount of NADH present by monitoring the magnitude of NADH fluorescence. (b) Fiber-optic transmission spectroscopy is used to measure the copper sulfate concentration in an electroplating bath. (c) Fiber-optic Raman spectroscopy is used to monitor the temperature and extent of curing in an industrial epoxy curing reaction (described in Section III.A.1). Raman spectroscopy is also used in the nuclear industry for detecting water in a sodium nitrate slurry and also has potential applications in the power industry such as on-line monitoring of boiler water chemistry, on-line monitoring of corrosion and deposits, and *in situ* inspection of steam generators during outages. The second approach is based on correlating spectral characteristics found over a range of wavelengths with the parameter of interest. Typically, an entire spectrum, such as an absorption or fluorescence spectrum, is collected and information is extracted from each spectrum by a suitable data-processing algorithm. This technique is used in the agriculture industry to determine parameters such as protein, water, and carbohydrate levels in grains.

In general, all the fiber-optic pH sensors as well as CO_2 and NH_3 sensors can be used for monitoring and for industrial process control. In particular, fiber-optic pH sensors for measuring acidity and alkalinity are employed in industries such as manufacturing, photographic developers, and waste treatment. A fiber-optic CO_2 sensor described above based on the pH-sensing mechanism has been used for on-line fermentation monitoring. It was also demonstrated that imaging fiber-optic pH sensors could be used to monitor pH changes continuously during beer fermentation and to monitor localized corrosion.

V. RECENT DEVELOPMENTS

Over the past several years, fiber-optic chemical sensors have received increasing attention because promising new technologies have been developed. New fiber-optic chemical sensor systems are based on integrating technologies from several different fields and disciplines including optics, chemistry, biology, and mechanical, electrical, and computer engineering. Recent advances in these fields have supported the development of improved fiber-optic chemical sensors with capabilities that are superior to current analytical methods. Technologies that influence fiber-optic chemical sensor development include (a) new data acquisition and data analysis software, (b) improved technologies for the production and design of new sensing materials through biological (e.g., recombinant DNA technologies) and chemical (e.g., molecular imprinting) approaches, and (c) development of new materials. In this section novel fiber-optic chemical sensors systems based on these new and advanced technologies are described.

A. Multianalyte Sensing

A key strength of optical fibers is the high information capacity they carry (high bandwidth). The high bandwidth capabilities of optical fibers are extensively employed for telecommunication applications and more recently for fiber-optic chemical sensors. Many different wavelengths can propagate through the fiber simultaneously allowing the transmission of multiple sensing signals arising from multiple analytes.

Multianalyte sensing is important for clinical, biological, environmental, and industrial analysis in which simultaneous detection of more than one analyte is required. For example, measurement of pH, O_2, CO_2, antibodies, DNA sequences, antibiotics, viruses, and bacteria in single blood samples can provide physicians with rapid and comprehensive information about a patient's medical condition.

Several approaches have been described for multianalyte fiber-optic chemical sensor construction. One approach involves the use of direct spectroscopy where different wavelengths are transmitted through the optical fiber to the sample and the returning light signals are analyzed by advanced data analysis procedures (see also Section IV.C). Another approach, briefly described in Section IV.A, involves assembling several fibers each with a different immobilized indicator dye into one fiber bundle. Alternatively, imaging fibers can be used by immobilizing discrete sensing regions, each containing a different indicator, at a precise location on the fiber's distal end (Fig. 20). A CCD detector is used to spatially resolve the signal obtained from each sensing element. In this way, as shown in Fig. 20, multiple analytes (O_2, pH, and CO_2) are monitored using a single imaging optical fiber. Furthermore, the use of an imaging fiber allows combined imaging and chemical sensing as will be described later in this section.

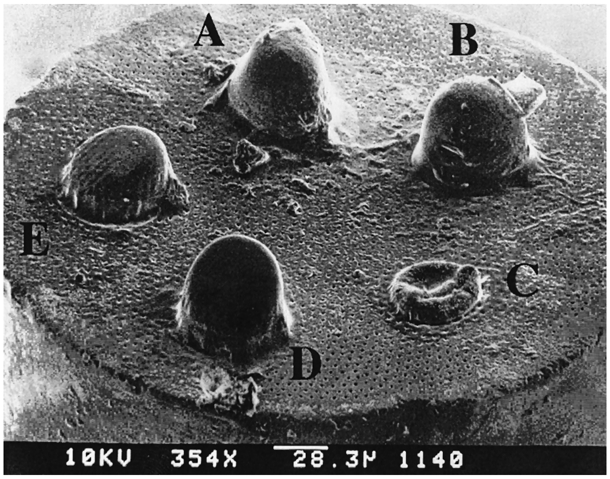


FIGURE 20 Multianalyte fiber-optic chemical sensor with different indicators immobilized in polymers attached to an imaging fiber. CO_2-sensitive matrices (A and B), pH-sensitive matrix (C), and O_2-sensitive matrices (D and E). [Reprinted with permission from Ferguson, J. A., Healey, B. G., Bronk, K. S., Barnard, S. M., and Walt, D. R. (1997). *Anal. Chim. Acta* **340,** 123–131.]

A more recent approach for multianalyte sensing with imaging fibers is based on the ability of each individual fiber to carry its own light signal. Thus, by attaching a sensing material to each individual fiber, an array of thousands of sensing elements can be constructed on an imaging fiber. In practice, the sensing elements are prepared by immobilizing fluorescent indicators on microsphere surfaces. The microspheres are distributed on the end of an imaging fiber containing thousands of microwells. These microwells are fabricated by selectively etching the individual cores on the tip of the imaging fiber as shown in Fig. 21. Since different indicators are immobilized on different microspheres, the array can be used to monitor multiple analytes. A CCD detector is used to monitor and spatially resolve the signal obtained from each microsphere. Imaging and data analysis software are used to calculate the analyte concentrations.

Multianalyte fiber-optic chemical sensors are in the first stages of research and development. Due to their importance for many analytical applications, it is expected that research efforts will continue to advance the capabilities of such sensors.

B. Distributed Chemical Sensing

Distributed optical fiber sensors are capable of spatially monitoring analyte concentrations along an optical fiber. Sensing elements are deposited longitudinally along the optical fiber and the signals obtained from each sensing element can be localized to a specific position along the fiber using a temporal feature of the signal. Distributed optical fiber sensors are needed in environmental applications such as detecting pollutant concentrations at different water depths (e.g., rivers and lakes). Distributed optical fiber sensors can identify a layer of sediment or water depth where the pollutant concentration is higher and thus provide valuable information about the cleaning strategy to be employed. Distributed optical fiber sensors are also useful for industrial applications where monitoring analyte gradients inside chemical or biological reactors may provide information about the process efficiency.

The distributed signals can be measured and spatially resolved by an optical time-domain reflectometry technique. An optical time domain reflectometer is based on the measurement of backscattered light attained from a light pulse propagating through an optical fiber. Light is backscattered because of inhomogeneities and impurities

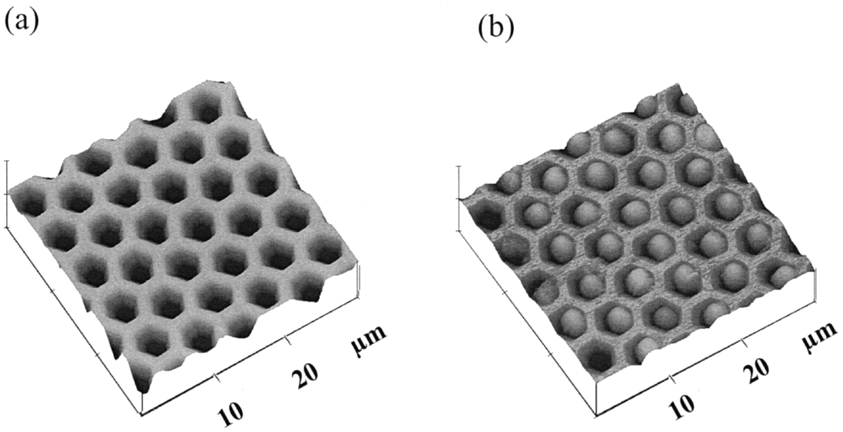


FIGURE 21 Scanning force micrograph (SFM) of a sensing microsphere array on an imaging fiber. (a) The microwells are fabricated by selectively etching the cores of the individual fibers composing the imaging fiber. (b) The sensing microspheres are distributed in the microwell.

in the silica comprising the optical fibers. The backscattered light signals are time-resolved and decrease as a function of distance. Changes in the backscattered light signals can be induced by sensing materials in response to changes in the concentrations of chemical species.

The evanescent field can be employed for sensing in distributed optical fiber sensors by measuring the analyte interactions at different positions along the fiber. Typically, the optical fiber cladding is removed at several points along the fiber and replaced by a thin polymer layer containing an indicator dye. For example, if a fluorescent pH indicator dye is used, by transmitting excitation light through the fiber in time intervals (usually a few nanoseconds) and measuring the backscattered emission, it is possible to observe changes occurring in each sensing element and thereby spatially resolve pH along the fiber.

Another method employed for distributed optical fiber sensors is based on microbend sensing. In this approach, the sensing elements are made by covering small areas along the fiber with hydrogels. The hydrogel can swell or shrink depending on the particular analytes present in the surrounding environment. The immobilized hydrogels' mechanical movement causes slight bending of the fiber. The bends in the fiber change the TIR conditions of light propagating through the fiber (see Fig. 3) resulting in a decrease in light transmission. By using optical time-domain reflectometry, the exact location and intensity of the bending can be determined and related to the spatial distribution of analytes in the sample. Based on the concept of microbending, and by using modified hydrogels that can respond to different analytes, various distributed optical fiber sensors can be constructed.

Although still in the research stage, it is expected that distributed sensing capabilities will improve many fiber-optic chemical sensor technologies. In particular, integration of distributed and multianalyte sensing will result in powerful fiber-optic chemical sensor devices.

C. Imaging and Chemical Sensing

Optical imaging fibers array can carry images from one end of the fiber to the other due to the coherent nature of the fibers. The imaging capabilities of such fibers are utilized simultaneously to image and measure local chemical concentrations with micrometer-scale resolution. In this technique, an imaging fiber's distal face is coated with analyte-sensitive materials, which produces a microsensor array capable of spatially resolving chemical concentrations. The concept is shown in Fig. 22. For example, a pH-sensitive array is fabricated by coating the imaging fibers with a pH-sensitive polymer layer containing a fluorescent dye. In this way, an optical sensor array is produced in which each pixel in the array imaging fiber is coated by a pH-sensitive layer and acts as its own individual sensor. This pH-sensitive sensor array can be used for both visualizing remote localized corrosion at metal

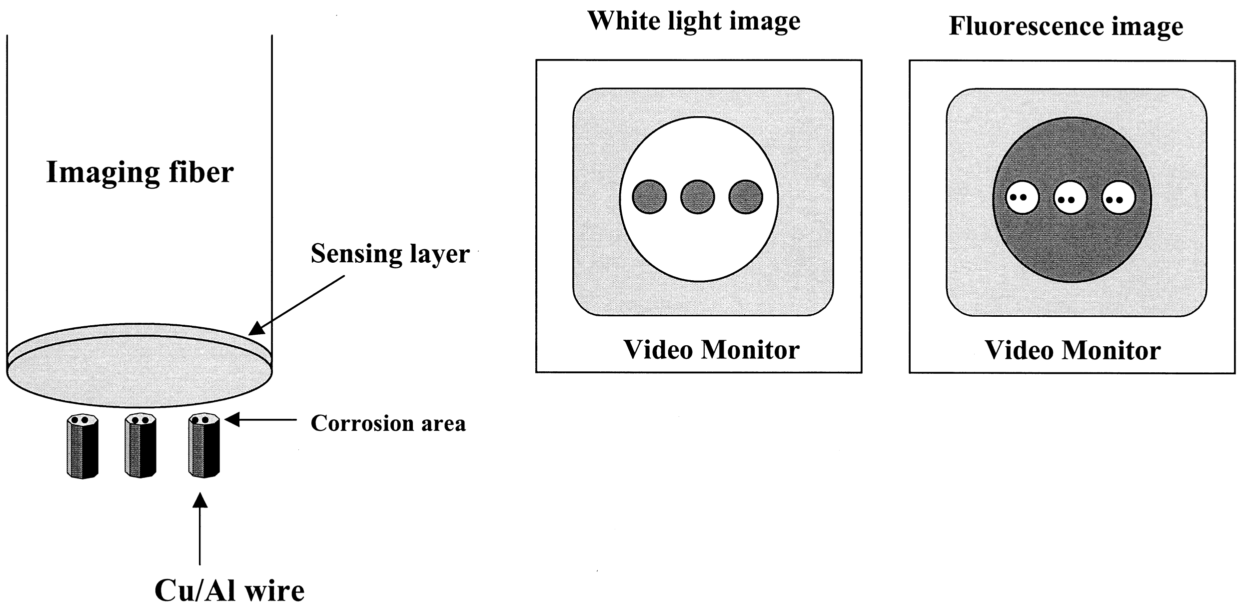


FIGURE 22 Combined imaging and chemical sensing concept. The technique provides the ability to both view a sample and measure surface chemical changes using a single optical imaging fiber.

surfaces and measuring local chemical concentrations. Corrosion monitoring of a copper/aluminum galvanic pair has been demonstrated. Cathodic and anodic reactions change the local pH. These processes were investigated with a pH-sensitive sensor array by measuring the pH-induced fluorescence changes occurring where the sensor contacted the metal surface. Such optical sensor arrays reduce the precision with which an extremely small probe must be positioned and offer major advantages over microelectrode arrays in both ease of fabrication and measurement on the micrometer scale.

VI. CONCLUSIONS

Fiber-optic chemical sensors offer several advantages over other sensing technologies based on the unique characteristics of optical fibers. The principal advantages include their immunity to harsh environmental conditions (e.g., electromagnetic interference, high temperature, high pH) and their ability to function without any direct electrical connection to the sample. These features have resulted in the development of different fiber-optic chemical sensors for analytical applications in the clinical, environmental, and industrial fields.

Recently, optical fibers have attracted attention mainly due to their use in telecommunications. New technologies have been developed for fabricating optical fibers with very efficient light transmission capabilities. Fibers can transmit very high amounts of information. In fiber-optic chemical sensors, this information can be different analytical signals resulting from different sensing elements located at the end of an optical fiber. It is expected that in the near future, advanced multianalyte fiber-optic chemical sensors will be able to accomplish real-time measurements for various analytical applications.

SEE ALSO THE FOLLOWING ARTICLES

Absorption • Analytical Chemistry • Biomaterials, Synthesis, Fabrication, and Applications • Environmental Measurements • Environmental Toxicology • Enzyme Mechanisms • Nucleic Acid Synthesis • Optical Fibers, Fabrication and Application • Optical Fiber Techniques for Medical Applications • Raman Spectroscopy

BIBLIOGRAPHY

Culshaw, B., and Dakin, J. (eds.). (1997). "Optical Fiber Sensors (Components and Subsystems)," Vol. 3, Artech House, Norwood, MA.

Dakin, J., and Culshaw, B. (eds.). (1997). "Optical Fiber Sensors (Applications, Analysis, and Future Trends)," Vol. 4, Artech House, Norwood, MA.

Lin, J. (2000). "Recent development and applications of optical and fiber-optic pH sensors," *Trends. Anal. Chem.* **19,** 541–552.

Mehrvar, M., Bis, C., Scharer, J. M., Moo-Young, M., and Luong, J. H. (2000). "Fiber-optic biosensors—trends and advances," *Anal. Sci.* **16,** 677–692.

Rogers, A. (1999). "Distributed optical-fiber sensing," *Meas. Sci. Technol.* **10,** R75–R99.

Rogers, K. R., and Poziomek, E. J. (1996). "Fiber optic sensors for environmental monitoring," *Chemosphere* **33,** 1151–1174.

Walt, R. D. (1998). "Fiber optic imaging sensors," *Acc. Chem. Res.* **31,** 267–278.

Wolfbeis, O. S. (2000). "Fiber-optic chemical sensors and biosensors," *Anal. Chem.* **72,** 81R–89R.

Hybridomas, Genetic Engineering of

Michael Butler
University of Manitoba

I. Introduction: The Nature of Antibodies
II. The Molecular Structure of Antibodies
III. Glycosylation of Antibodies
IV. Production of Monoclonal Antibodies
V. Immunization *in Vivo*
VI. Immunization *in Vitro*
VII. The Development of Cell Hybridization
VIII. Methods of Cell Fusion
IX. Cell Fusion To Immortalize Lymphocytes
X. Selectable Gene Markers for Cell Selection
XI. Clonal Selection of Mab-Secreting Hybridomas
XII. Assay of Monoclonal Antibodies
XIII. Human Monoclonal Antibodies
XIV. Recombinant Antibodies
XV. Recombinant Antibody Fragments
XVI. Therapeutic Antibodies
XVII. Antibodies from Plants
XVIII. Humanized Antibodies from Transgenic Mice
XIX. The Importance of Glycosylation to Therapeutic Antibodies
XX. Large-Scale Production of Monoclonal Antibodies from Hybridomas
XXI. The Control of Culture Parameters
XXII. Serum and Serum-Free Medium for Antibody Production from Hybridomas
XXIII. Conclusions

GLOSSARY

Cell fusion (or hybridization) The process of two cells fusing (or hybridizing) into one. The resulting fused cell is called a heterokaryon and might be quite unstable genetically.

Constant region The region of an immunoglobulin in which the amino acid sequence does not show changes within a given antibody isotype.

Electrofusion The application of electrical impulses to cause cell fusion.

Fusogen A substance that will cause the fusion of two cells.

Glycosylation The metabolic pathway found in eukaryotic cells that allows the addition of a carbohydrate group on to a protein.

Heterokaryon The product of the fusion of two cell types. The membrane systems of the two cells fuse and the cytoplasm combine prior to fusion of the genetic material. The initial fused cell may be quite unstable genetically.

Hybridoma A stable hybrid cell that secretes a

monoclonal antibody. The hybridoma is created from the fusion of an antibody-secreting B-lymphocyte and a transformed myeloma. The term was first used in the 1970s following the breakthrough work of Kohler and Milstein.

Immunoglobulin Proteins found in the blood that show antibody activity.

Monoclonal antibody An antibody that is specific to a single antigen. A monoclonal antibody is synthesized from a homogeneous population of hybridoma cells.

Polyethylene glycol This is a commonly used fusogen for the fusion of two cells.

Quadroma A cell formed by the fusion of two hybridomas. The immunoglobulin product of a quadroma will contain a mixture of heavy and light chain structures derived from each parental line.

Variable region The region of an immunoglobulin in which the amino acid sequence changes so that the molecule can bind to a specific antigen.

HYBRIDOMAS are hybrid cells capable of the continuous production of monoclonal antibodies. They combine the key properties of the two parental cells: a myeloma with an infinite life span and a B-lymphocyte capable of synthesizing a single antibody. The technology for producing hybridomas was developed by Kohler and Milstein who gained the Nobel Prize in 1984. Hybridomas can be grown in suspension in large bioreactors for the production of kilogram quantities of monoclonal antibodies. The antibodies have a range of applications because of their high specificity in recognizing selected proteins. This enables them to be used for diagnosis and testing in applications such as blood typing, the detection of virus, pregnancy testing or for the detection of contaminants in food. The application of monoclonal antibodies as human therapeutic agents in the treatment of disease has been suggested for a number of years. However, there have been difficulties in the production of antibodies that are not immunogenic to humans. In the late 1990s a range of human or "humanized" antibodies have been produced specifically for the treatment of cancer. The number of such therapeutic monoclonal antibodies is likely to increase in the future as a result of the numerous clinical trials that are now taking place.

I. INTRODUCTION: THE NATURE OF ANTIBODIES

Antibodies are glycoproteins found in body fluids including blood, milk, and mucous secretions and serve an essential role in the immune system that protects animals from infection or the cytotoxic effects of foreign compounds. Antibodies will bind with high affinity to an invasive molecule. Normally the binding is to only part of a large molecule (the epitope) and so there may be many different antibodies for a particular compound. Antibodies have become essential tools for biological research because of their very specific recognition and affinity for one compound (the antigen). This has not only led to the use of antibodies in the recognition of specific cellular components but also to the development of routine diagnostic medical tests. More recently antibodies have been used as therapeutic agents for the treatment of human disease.

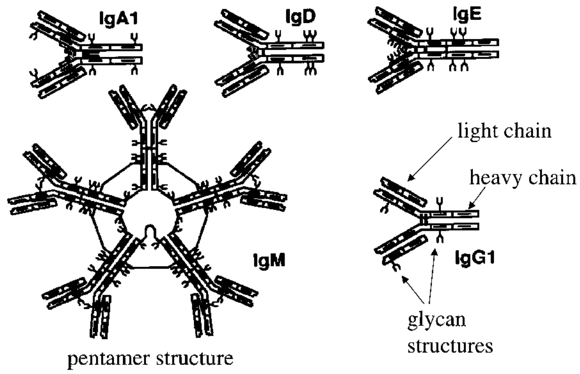


FIGURE 1 Structures of immunoglobulin isotypes.

Each B-lymphocyte is capable of producing one type of antibody in response to a particular antigen which interacts with a cell surface receptor. Stimulation by an antigen causes growth and an expansion of the cell population capable of producing the corresponding antibody. The variety of antibodies present in any animal reflects the population of B-lymphocytes which have been stimulated by previous exposure to a range of antigens.

Antibodies are found in a specific protein fraction of blood called the gamma-globulin or the immunoglobulin fraction. They are synthesized by a subset of white blood cells—the B-lymphocytes. The molecular structures of the five major classes (isotypes) of immunoglobulins (IgM, IgD, IgG, IgE, and IgA) are shown in Fig. 1. The basic structural arrangement of two heavy associated with two light chains is similar for all the isotypes. However, each isotype is distinguished by different heavy chain structures which are of varying length, number of domains, and glycan structures. The glycans are indicated by the fork structures (⊔). It is also to be noted that the IgM configuration consists of five basic structures linked as a pentamer.

II. THE MOLECULAR STRUCTURE OF ANTIBODIES

A structural representation of an antibody (immunoglobulin, IgG) which has an overall molecular mass of 150 kD is shown in Fig. 2. This is the major class of immunoglobulin found in blood serum. The molecular structure consists of two light and two heavy chains bound by disulfide

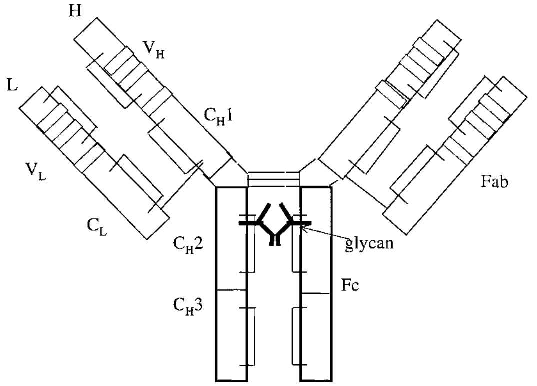


FIGURE 2 Antibody structure: IgG based on schematic diagram.

bridges. The heavy chain of IgG has four domains VH–CH1–CH2–CH3 and the light chain has two domains VL–CL. The "constant" region (C) of a particular immunoglobulin class varies only with the species of origin. For example, a human IgG would have a different constant region from a mouse IgG. The variable domains (V) account for the diversity of antibody structure. Digestion of the molecule with papain cleaves the heavy chain in the "hinge" region and results in three fragments. Two Fab (antibody-binding fragments) each contain the N-terminal end of a heavy chain with disulfide linked light chain. The other fragment is the Fc which consists of the C-terminal end of the two heavy chains. There are two glycan structures present in the space between the two CH2 domains. In some immunoglobulins there are also glycans present in the variable region of the molecule.

Figure 3 shows an alternative representation of an antibody structure based upon X-crystallographic data. Here the unique antigen-binding site which consists of hyper-variable sequences of amino acids is shown clearly. These are formed from three hypervariable loops (complementarity determining regions, CDR) of the VH domain and three hypervariable loops of the VL domain. The variable sequence is produced by somatic recombination and by mutagenesis and accounts for the diversity of antibody molecules. This region enables the antibody to bind to one

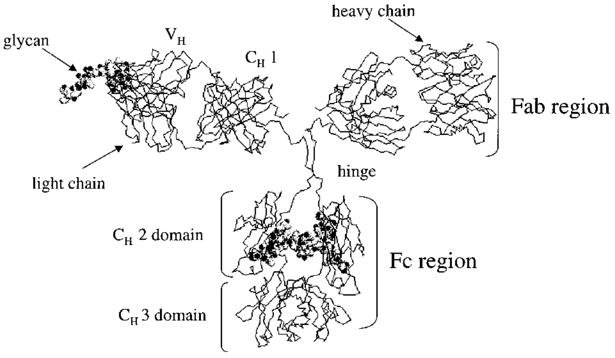


FIGURE 3 Antibody structure—IgG based on X-ray crystallography data.

specific molecule (called the antigen) with high affinity. The other important functional component of the molecule is the effector site which is found in the constant region. The effector functions can be mediated by binding complement (C1q) and those mediated by binding to Fc receptors of specific cells. Complement activation leads to the activation of leukocytes and phagocytosis. The Fc receptors are on certain cells of the immune system such as phagocytes and natural killer (NK) cells. Binding to receptors in these cells produces a variety of biological responses including antibody-dependent cellular cytotoxicity (ADCC), phagocytosis, endocytosis, and release of inflammatory agents.

A particular antibody will be produced in an animal following the injection of the corresponding antigen. For example, if human insulin is injected into a mouse, after a few days the blood will contain significant quantities of mouse antibody capable of binding to human insulin. The immunoglobulin fraction of the mouse blood can be extracted and will contain the required anti-insulin. However, this fraction will also contain numerous other antibodies and it would be very difficult to isolate the particular antibody that may be required. Because of the multiplicity of immunoglobulin types in the fraction the term "polyclonal antibody" is used. This polyclonal antibody may even include different antibodies against insulin. These would be antibodies reactive to different regions (epitopes) of the insulin molecule.

III. GLYCOSYLATION OF ANTIBODIES

Antibodies are glycoproteins containing variable glycan structures. A single conserved N-glycan site is contained in IgG on each CH2 domain of the Fc region at Asn-297 (Fig. 4). The carbohydrate is of a biantennary complex type. The structural variability is associated with a bisected GlcNAc (+/−), core fucosylation (+/−), non-, mono-, or digalactosylation and possible sialylation. The glycosylation of the Fc region is essential for effector functions of the antibody such as complement binding, binding to Fc receptors, induction of antibody-dependent cytotoxicity (ADCC), and the half-life in the circulatory system. Around 20% of human antibodies are also glycosylated in variable region of the Fab fragment. This glycan may

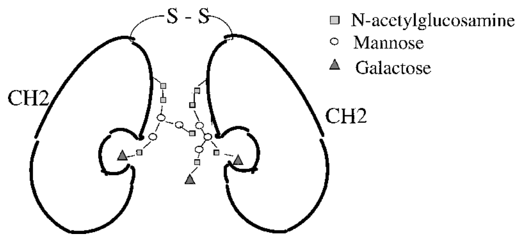


FIGURE 4 Glycan interaction on CH2 domain of IgG.

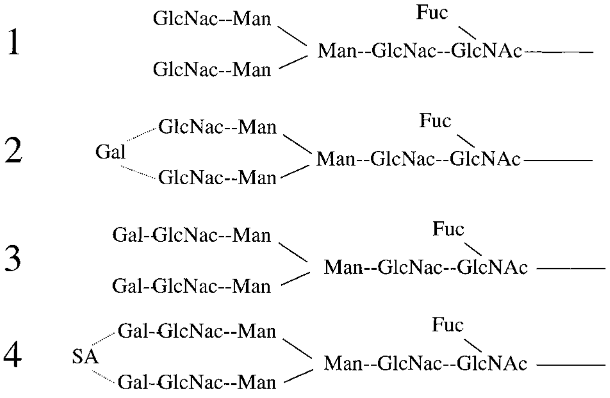


FIGURE 5 Glycoforms of IgG.

be important for antigen binding with specific examples showing that the degree of binding may either increase or decrease.

Although the level of glycosylation of IgG is small (2–3% by weight) compared to other proteins, the glycan structures on immunoglobulins are known to have a significant effect on immune responses. Figure 5 shows the common glycoforms of IgG with 0, 1, or 2 galactose terminal residues (G0, G1, and G2) and the possibility of a sialic acid terminal group on G2. The distribution of these glycoforms changes under certain pathological conditions. For example, it is well established that a high proportion of agalactosylated glycan structures in immunoglobulin (G0) is associated with specific human disorders, notably rheumatoid arthritis. Here the predominant form of the glycan attached to Asn-297 is a biantennary complex structure that terminates in N-acetylglucosamine residues and lacks the usual galactose terminus. This altered glycan structure results in a change in the interaction of the immunoglobulin with specific monocyte receptors. Also, there are changes to the structure of the immunoglobulin caused by the altered glycan that may explain changes in immune response related to the disease condition.

The glycosylation of IgM is more complex with five identifiable N-glycan sites in the heavy chain. Three of these are complex biantennary, whereas the other two are of a high mannose type glycan. Variations of these glycan structures may also produce undesirable immune responses if they are utilized as therapeutic products of hybridomas.

IV. PRODUCTION OF MONOCLONAL ANTIBODIES

This section describes the background to the methods used to produce hybridomas that secrete monoclonal antibodies (Mabs). This is a technology that developed as a laboratory technique but now can be performed on a large commercial scale in bioreactors. The techniques described can be performed in any cell culture laboratory with access to animal facilities. Alternatively, a wide range of antibody-secreting hybridomas are available from culture collections.

In the 1970s the techniques of immunization, cell hybridization, genetic selection, and cell cloning were utilized to produce cells that were hybrids of B-lymphocytes and myelomas. The B-lymphocytes are antibody-secreting cells, whereas the myeloma cells are transformed lymphocytes capable of growing indefinitely. The resulting hybridomas were capable of continuous synthesis of preselected antibodies. Kohler and Milstein obtained the Nobel Prize in 1984 for their work on the development of Mab-secreting hybridomas. The original work described the creation of a mouse–mouse hybridoma that secreted antibody with affinity to sheep red blood cells as antigens. The antibody could be easily detected by a hemolytic plaque assay that showed the capability of the antibody to bind to and lyse sheep erythrocytes. Since the 1970s, these methods have found wide application and have resulted in the large-scale production of kilogram quantities of some monoclonal antibodies. The term "monoclonal" indicates that the antibody is of a single type. This will bind to just one antigen.

The method developed by Kohler and Milstein involves four stages which result in the production of a hybrid lymphocyte with an infinite growth capacity and capable of continuous synthesis of a single antibody. The stages of this process are summarized in Fig. 6. The four stages involve

- Immunization
- Cell fusion
- Genetic selection
- Cell cloning

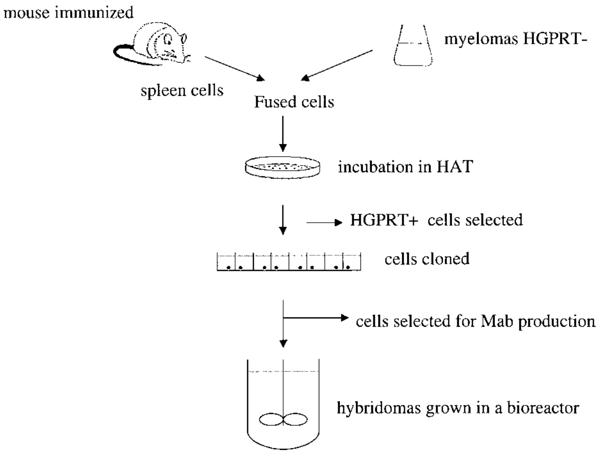


FIGURE 6 Production of Mab-secreting hybridomas.

V. IMMUNIZATION *IN VIVO*

The first stage of the production of a hybridoma is to obtain lymphocytes from an animal that are enriched with specific antibody secreting cells. Immunization involves the injection of a chosen antigen into an animal—mice and rats have been most commonly used. The time required to produce an immune response resulting in antibody synthesis will depend upon the antigen but a period of up to 3 to 4 weeks ensures maximum response. Large molecules tend to produce a strong response over a short period of time. Small molecules are often conjugated to carrier proteins such as albumin and multiple injections spaced over several days may be necessary to enhance the immune response.

Antibodies are synthesized by B-lymphocytes which can be isolated from the spleen of an immunized animal. The isolated spleen is homogenized gently and the lymphocyte cell fraction collected by centrifugation. Approximately 1% of the cell population isolated from the spleen will secrete antibodies. At this stage the cell fraction is a mixed population with a limited capacity for growth.

VI. IMMUNIZATION *IN VITRO*

Although most laboratories use mice for producing active lymphocytes, an alternative approach involves immunization *in vitro*. This process involves the activation of cells obtained from the spleen of a non-immunized mouse. The cells should be suspended in a medium containing the selected antigen along with various stimulating growth and differentiation factors. These factors can be supplied from culture medium following incubation with mixed lymphocytes (or thymocytes). This is called "conditioned" medium and contains various growth-promoting factors secreted by the cells. These factors are called cytokines and include interleukins, B-cell growth factor and B-cell differentiation factor. Some of these cytokines have now been well defined and are available as recombinant proteins from commercial suppliers. The effectiveness of immunization *in vitro* is dependent upon the optimal combination of these factors during cell activation.

An advantage of immunization *in vitro* is that the activation of B-lymphocytes takes 3 to 4 days rather than a few weeks as is the case with immunization *in vivo*. Furthermore, weak antigens at low concentrations can be used. A disadvantage is that certain immunoglobulin isotypes tend to be produced preferentially (usually IgM) although refinement of the techniques during cell activation can stimulate the production of other isotypes.

VII. THE DEVELOPMENT OF CELL HYBRIDIZATION

In 1965 Harris and Watkins reported that inactivated Sendai virus caused the hybridization of a mixed population of human HeLa cells and mouse Ehrlich ascites tumor cells. The result of the fusion was a mixed population of hybrid cells (called heterokaryons) that were genetically unstable. Figure 7 indicates the sequence of events during fusion showing the cytoplasmic fusion of two dissimilar cells followed by the hybridization of the nuclei of the two cells. After a period of growth the heterokaryons tended to lose some of their genetic material and become stable hybrids retaining some of the phenotypic characteristics of each parental cell. The method turned out to be an extremely valuable tool for biological research. In 1969 Harris showed that when normal cells were fused with malignant cells the malignant phenotype was not always retained. This was the first direct evidence for the existence of human suppressor genes, derived from the normal cells and that could result in suppression of the tumorigenic characteristics. These genes whose products include the retinoblastoma protein and p53 are now well characterized in terms of their role in malignancies.

The cell hybridization technique has also been useful in developing an understanding of cell differentiation and gene regulation. For example, the normally quiescent genetic material of highly differentiated cells can be reactivated following fusion with cells actively engaged in protein synthesis. This was shown by Harris in the late 1960s by the fusion of a cell population of chicken erythrocytes and growing HeLa cells.

Cell fusion has also been used extensively in human chromosome mapping. The heterokaryons resulting from the fusion of human and mouse cells are genetically unstable and tend to lose human chromosomes randomly. This eventually gives rise to a mixed population of stable hybrids from which individual cell clones can be isolated. Many of these clones may contain single human chromosomes. It is the association of a particular chromosome in an isolated cell clone with a selected measurable phenotypic characteristic such as an enzyme activity that

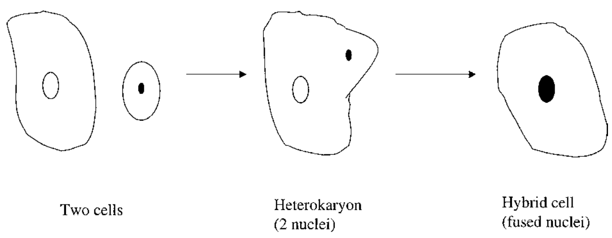


FIGURE 7 Cell hybridization.

allows the gene of that enzyme to be assigned to a specific human chromosome. With the use of this technique, many human genes have been assigned to particular chromosomes (known as "chromosome mapping").

It was this same technique of cell fusion that Kohler and Milstein used in their work reported in 1975 that allowed the creation of stable hybrid cells from the hybridization of antibody-secreting B-lymphocytes with transformed myelomas. The resulting cells retained two important phenotypic characteristics from the parents—the ability for infinite growth (from the myeloma) and the ability to synthesize antibody (from the lymphocyte). The original objective of this work was to study somatic mutation as a mechanism for antibody diversity. This is the ability of B-lymphocyte to go through a maturation process following initial contact with an antigen to produce antibodies of increasing affinity. However, the application of the cell fusion technique to produce antibody-producing cells with an infinite growth capacity had a major impact on the ability to produce large quantities of antibodies that could be used for a variety of functions both in biological research and also as medically important products. The term "hybridoma" was derived in 1976 by Herzenberg and Milstein to describe a homogeneous clone of these antibody-producing hybrid cells. The term "monoclonal antibody" refers to the secreted product of the cells. Unlike antibodies derived from blood samples ("polyclonal"), the monoclonal antibodies from a single hybridoma are molecularly homogeneous and have a specific affinity for a particular antigen.

VIII. METHODS OF CELL FUSION

Cells can be induced to fuse if two cell populations are brought close together at a high cell concentration (10^6 to 10^7 cells per well of a 96 multi-well plate) in the presence of viruses or by chemical agents (called "fusogens"). The process involves a destabilization of adjacent cell membranes which eventually fuse to form a hybrid cell. Initially, two distinct nuclei are present in the fused cell (a heterokaryon). Eventually the nuclei fuse to produce a stable hybrid cell.

Although UV-inactivated Sendai viruses were originally used as agents for cell fusion, the more widely used method is now fusion by the chemical agent polyethylene glycol (PEG). This is a polymer, available at a molecular weight range of 200–20,000 kD. PEG at 4000–6000 kD is most suitable for cell fusion. Cell fusion can occur in a solution of PEG (40–50% w/v) within 1–2 min. In this process, cell swelling accompanies fusion. This enables adjacent cells to approach very closely and also the plasma membrane becomes permeable to small ions. However, lysis of swollen cells may also occur.

Alternatively, electrofusion can be used. In this technique, two populations of cells are introduced into a small sterile chamber. An electric current is applied in high-voltage pulses for short time periods. During this period the membrane will become highly permeable. This is similar to the process of electroporation used to facilitate the entry of DNA into cells. This causes the cells to orientate along the line of the current and fuse. This process is highly efficient, producing a high percentage of viable hybrid cells. The most suitable voltage for electrofusion is one that causes approximately 50% death in the cell population. This would typically be around a voltage of 200 V for a cell pellet held in a small electroporation cuvette. From such a protocol there may be around 50 fused cells from an original total of 5×10^6 cells from each parental cell line.

IX. CELL FUSION TO IMMORTALIZE LYMPHOCYTES

Although immunization can result in lymphocytes capable of producing the required antibody, the cells will only grow for a limited period of time. The purpose of lymphocyte hybridization is to combine the desired property of antibody synthesis of the B-lymphocyte population with the infinite growth capacity of a myeloma. Therefore, the selected lymphocytes are fused with a population of myeloma cells. Those commonly used for mouse or rat cells are shown in Table I. Suitable myeloma fusion partners are selected for two other important characteristics:

- Nonproduction of antibodies. This is desirable so that the resulting hybridoma does not synthesize more than one antibody.
- Possession of a genetic marker, such as the lack of an enzyme, to allow cell selection. For example, myelomas deficient in HGPRT (hypoxanthine guanine phosphoribosyl transferase) are commonly used. This allows selection of hybridomas in HAT medium (see "selectable gene markers").

TABLE I Rodent Cell Lines (Myelomas) Commonly Used as Fusion Partners

Species	Cell line	Immunoglobulin expression
Mouse	X653	No
	NS0	No
	Sp2/0	No
	NS1	Yes
Rat	YB2/0	No
	Y3-Ag	Yes

To allow fusion the activated B-lymphocytes are mixed with a suitable fusion partner (the myeloma) in a medium containing polyethylene glycol (PEG). A proportion of the cells will fuse within a minute under these conditions.

X. SELECTABLE GENE MARKERS FOR CELL SELECTION

The process of cell fusion will result in a heterogeneous population of cells that will contain unfused parental cells, lysed cells as well as the required hybrid cells. At this stage cell selection is important so that the hybrid cells can be isolated from the mixture. For hybridomas there are two important stages of cell selection:

- Isolation of hybrid cells from parental cells
- Selection of antibody secreting cells within the hybrid cell population.

The basis of cell selection is to distinguish cell types through genetic differences. One required characteristic of the resulting hybridomas is the ability for effective growth in culture. So, initial cell selection can involve the incubation of the mixed population of cells in a suitable culture environment. This includes the addition of a suitable liquid growth medium to the cells and incubation at 37°C for a few days. This will allow the growth of all viable cells which will include the hybridomas and myelomas. Growth of these cells will dilute any nonviable and lysed cells from the mixture. Unfused lymphocytes have only a limited capacity for growth and will also be eventually eliminated.

The hybridomas are selected from the myelomas on the basis of a genetic marker that is normally applied to the myelomas. The most commonly used genetic marker is $HGPRT^-$ which indicates a cell with a defective enzyme, hypoxanthine guanine phosphoribosyl transferase. This is a normal metabolic enzyme that is capable of catalyzing the addition of phosphoribosyl pyrophosphate on to hypoxanthine or guanine. This constitutes the salvage pathway for converting purines to nucleotides as part of nucleotide synthesis in all normal cells (Fig. 8). However, there is an alternative pathway of nucleotide formation which involves the complete synthesis of the purine ring (called the "*de novo*" pathway) from smaller metabolic precursors. Normal cells can utilize either pathway depending upon the nutrient precursors available in the surrounding media.

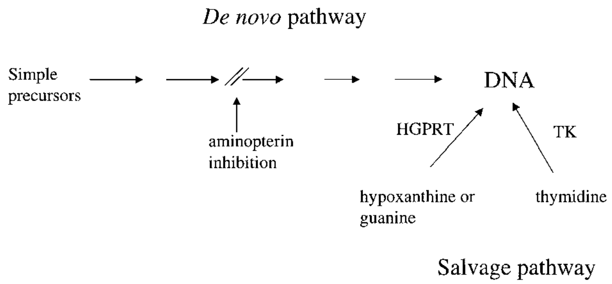


FIGURE 8 *De novo* and salvage pathways of nucleotide synthesis.

The myelomas chosen for cell hybridization with lymphocytes have an $HGPRT^-$ genetic marker through previous random mutation and selection. However, the hybridomas would be expected to have normal HGPRT activity ($HGPRT^+$) because they would receive the normal gene from the parental lymphocyte. Either cell line would be able to grow normally in standard cell culture medium in which most nutrients are provided. The *de novo* pathway for nucleotide synthesis would operate in both cells, although the salvage pathway would only occur in the hybridomas.

The principle of selection is to place these cells into a "selective medium" in which the *de novo* pathway of nucleotide synthesis is inhibited. The key to this is the compound, aminopterin which is an analogue of folic acid and a specific inhibitor of dihydrofolate reductase, an essential enzyme for the formation of tetrahydrofolate (FH_4) required as a coenzyme of the *de novo* purine nucleotide synthesis pathway. Tetrahydrofolate is also required for the formation of thymidine. However, if hypoxanthine and thymidine are provided in the culture media of $HGPRT^+$ cells they will be able to grow normally. On the other hand $HGPRT^-$ cells would have no means of synthesizing purine nucleotides and consequently would be unable to grow.

The selective medium, HAT contains hypoxanthine, aminopterin, and thymidine. This medium allows the selection and growth of hybridomas which are $HGPRT^+$ and have a normal functioning salvage pathway. However, HAT is unable to support the growth of the $HGPRT^-$ myelomas because the *de novo* pathway is inhibited and the salvage pathway cannot function because of the defective enzyme.

In the mixed population of cells resulting from fusion, the newly formed hybridomas will be $HGPRT^+$ by inheritance from normal lymphocytes whereas the unfused myelomas carry the mutant $HGPRT^-$ marker. Therefore incubation in HAT will allow survival of the $HGPRT^+$ hybridomas but not the $HGPRT^-$ myelomas. So, after a few days incubation in HAT the culture will contain only hybridomas.

XI. CLONAL SELECTION OF Mab-SECRETING HYBRIDOMAS

After genetic selection with HAT the culture contains hybridomas but only some of these will secrete antibodies. Although the efficiencies of synthesis vary considerably, about 10% of the population of hybridomas formed from

cell fusion should be expected to secrete antibody. The next stage involves selection of Mab-secreting hybridomas from the population which has survived HAT treatment. Cell clones can be isolated by the method of limiting dilution. Cloning ensures that all cells in the cultures are genetically identical. This involves dispensing a cell suspension into a 96-well plate, so that each well contains an average of one cell. Hybridomas grow poorly at low densities but growth can be supported by feeder layer of cells. This consists of a population of cells (e.g., thymocytes, macrophages or splenocytes) which has been gamma irradiated to prevent any growth. However, the metabolism continues and secreted growth factors can help the growth of viable hybridoma cells particularly if they are inoculated at low density. Feeder layers can be purchased as frozen suspensions.

After allowing 1–2 weeks for growth, the medium of each well should be tested for antibody content using a suitable assay. Wells containing a high antibody titer should then be selected for further cell growth. At this stage the cells may be genetically unstable and a second round of cloning is recommended to ensure the isolation of a stable population of high-level antibody-secreting cells. This involves further testing of the culture medium for antibody content.

XII. ASSAY OF MONOCLONAL ANTIBODIES

There are three assay procedures that are commonly used to detect monoclonal antibodies in solution. Each is suitable for the measurement of Mab concentration in culture media.

- ELISA—enzyme-linked immunosorbent assay
- RIA—radioimmunoassay
- Affinity chromatography

ELISA is the most commonly used assay for antibodies and is adapted to multi-well plates for analyzing multiple samples. RIA is more sensitive but is more time consuming and expensive. Affinity chromatography is ideal if a high-performance liquid chromatography (HPLC) system is available and the hybridomas are growing in a serum-free medium. The basis of the three types of assay are described here:

A. Enzyme-Linked Immunosorbent Assay (ELISA)

This is a solid-phase binding assay that can easily be performed in a 96-well plate (Fig. 9). ELISA measures antigen or antibody concentration, depending on the protocol used. The stages of a typical assay involve a series of additions in which each component binds to the one previously added:

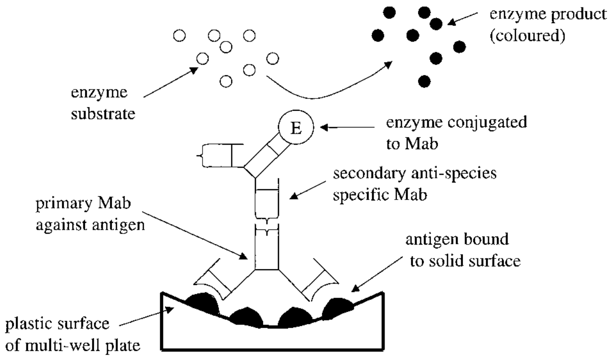


FIGURE 9 Enzyme-linked immunosorbent assay: ELISA.

- An appropriate antigen is bound to the plastic of the base of the plate. Most large proteins bind spontaneously. Difficulties with binding of small molecular antigens can be overcome by forming a conjugate with a larger molecule such as bovine serum albumin (BSA).
- The remaining attachment sites are blocked on the solid support by addition of a noninterfering protein such as BSA.
- A solution of the Mab under test or a standard antibody solution is added. This will bind to the antigen held on the solid support.
- An antibody–enzyme conjugate with specificity against the first antibody is added. The second antibody is species specific. This means that it binds to any immunoglobulin of the species from which the first antibody is derived, e.g., goat anti-mouse antibody. Conjugated to the second antibody is an enzyme such as alkaline phosphatase which can be detected by a colorimetric assay. These conjugated anti-Ig antibodies are available commercially.
- Finally a suitable enzyme substrate is added. The substrate is one which can be changed to a colored product by the enzyme bound to the conjugate. For example, p-nitrophenyl phosphate would be suitable for an alkaline phosphatase conjugate. The extent of coloration in each well of a plate can be measured by a multi-well reader.

B. Radioimmunoassay (RIA)

A solid-phase binding assay similar to ELISA can be adapted to radioactive detection when a radioactively labeled antigen or antibody is used. The radioimmunoassay (RIA) is more sensitive and reliable than ELISA but is usually more time consuming and more expensive because of the cost of the radioactive label.

C. Affinity Binding

Certain bacterial cell wall proteins (called Protein A and Protein G) bind with high affinity to mammalian immunoglobulins. Protein A is derived from *Staphylococcus aureus* and has a strong affinity for antibodies. This allows antibodies to be isolated by chromatography columns which contain inert beads conjugated to Protein A or G. If a sample (such as Mab-containing culture medium) is run through the column at neutral pH only antibodies will bind, allowing all other components to be washed out. Pure antibody can then be eluted from the column by a low pH buffer.

Suitable affinity columns of this type have been designed for use with HPLC and this offers an extremely rapid method of analyzing or purifying antibodies. However, the method will detect any mammalian immunoglobulin which means that the immunoglobulin content of the serum used in the growth medium may interfere with analysis.

HPLC affinity columns (such as ProAnaMabs from Hyclone) offer a rapid assay for measuring antibodies in serum-free culture medium and they could be used instead of ELISA. Affinity chromatography can also be used for large-scale antibody extraction, although the preparative Protein A or G columns are expensive.

XIII. HUMAN MONOCLONAL ANTIBODIES

Although murine-derived monoclonal antibodies are widely used as laboratory reagents, in affinity purification and clinical diagnostic tests, they have had limited success in human therapy. Immunoglobulins synthesized from mice and humans have different constant regions and so any antibody of mouse origin injected into a human could elicit an undesirable immune reactions. First, although the antigen-binding site might be appropriate for the target the antibody will not produce appropriate human effector responses such as those of complement and Fc receptor binding. Second, the human immune system will produce antibodies against the murine immunoglobulin. This is referred to as the human anti-murine antibody immune response (HAMA).

This presents an obstacle in developing therapeutic antibodies from a murine source. However, there are at least three major difficulties in producing human hybridoma cells capable of secreting human monoclonal antibodies.

A. The Source of Antibody-Secreting Lymphocytes

In generating murine hybridomas, the spleen of an immunized mouse is used as a source of the mixed lymphocyte population for cell selection. Clearly this is not possible with humans and the source of human lymphocytes is limited to samples of peripheral blood. These can be taken from patients who have acquired an immunity against a particular compound or disease. Alternatively, methods of *in vitro* immunization of human lymphocytes are possible. This approach requires the optimization of conditions for human B-lymphocyte activation by use of the appropriate cytokines and growth factors.

TABLE II Human or Human Hybrid Fusion Partners

Cell type	Cell line	Immunoglobulin expression
Human myeloma	SK007	Yes
Human lymphoblastoid	RH-L4	Yes but nonsecretor
Human lymphoblastoid	GM1500	Yes
Human lymphoblastoid	KR4	Yes
Human lymphoblastoid	LICR-LON	Yes
Human/human hybridoma	KR12	Yes
Human/mouse hybridoma	SHM-D33	Yes but nonsecretor

B. Immortalization and Chromosome Instability

There must be a suitable human fusion partner to immortalize the B-lymphocytes. Human myeloma cell lines are difficult to grow in culture. Human lymphoblastoid cell lines have been used as fusion partners but the frequency of cell fusion and genetic stability of the resulting hybridomas is low compared with equivalent fusions with mouse cells. An alternative approach is to immortalize the activated human lymphocytes by transfection with oncogenic DNA or by transformation by a virus.

C. Antibody Secretion of Human Parental Fusion Partners

Mouse myelomas commonly used in fusion are nonantibody secretors. The value of this is that the resulting hybridomas only secrete the antibody associated with the fused B-lymphocyte. Therefore, the culture product will be a single selected antibody type. However, most of the human myeloma or lymphoblastoid cells commonly used for hybridization are immunoglobulin secretors (Table II). This means that the resulting selected human hybridoma will secrete at least two antibodies, which are those associated with each of the parental cells.

XIV. RECOMBINANT ANTIBODIES

A further possibility is the humanization of monoclonal antibodies originally produced from mice. This process involves antibody engineering which relies on the

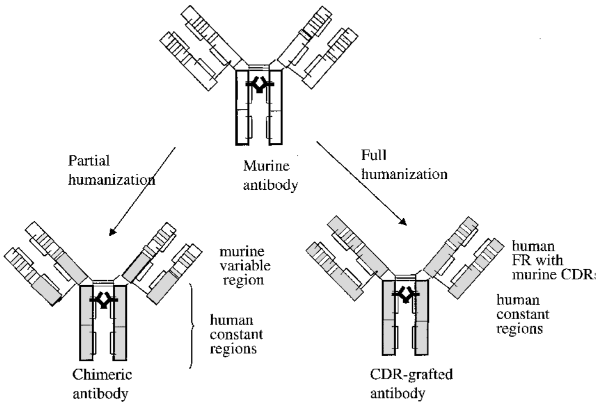


FIGURE 10 Ways to humanize an antibody.

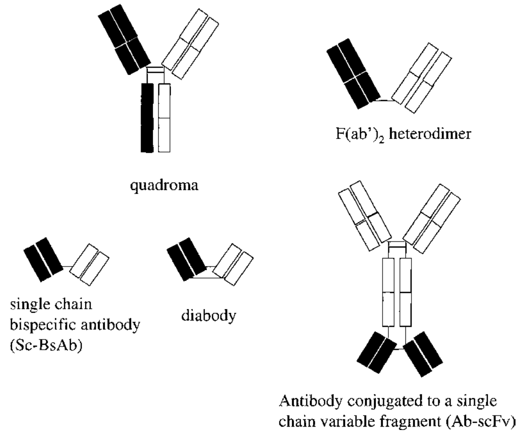


FIGURE 11 Antibody configurations.

techniques of recombinant DNA technology to rearrange some of the molecular domains of an immunoglobulin. Examples of these are shown in Fig. 10. In a chimeric antibody the mouse variable regions are linked to human constant regions. Thus in such a construct the antigen-binding site of the murine antibody is retained but the human constant region contributes the immunogenicity through the effector functions. A further step to humanizing the antibody by replacing portions of the V region that are not required for the antigen-binding site. The framework regions (FR residues) which were originally murine are replaced by human regions. Thus only the complementarity-determining regions (CDR) are retained as of murine origin. Hybrid antibodies of this type have now been used as human therapeutic agents.

The elimination of the murine constant regions reduces the previously experienced HAMA response. It is not always certain that complete humanization has an advantage over a chimeric antibody because humanization of the V region may result in a loss in affinity to the antigen. Also, it is not clear that the problem of unwanted immunogenicity can be totally removed because repeated doses of even a fully humanized antibody may elicit an anti-idiotype response, that is directed against the antigen-binding site. However, these developments in humanized therapeutic antibodies have allowed the introduction of a range of products against specific human diseases.

XV. RECOMBINANT ANTIBODY FRAGMENTS

Various fragments of human immunoglobulins have been expressed successfully in bacterial cells. These include the Fv fragment, the single-chain Fv fragment (scFv), the Fab fragment and the F(ab)2 fragment (Fig. 11). The Fv is the smallest antigen-binding fragment of an immunoglobulin with a molecular mass of around 25 kD. The VH and VL domains of the Fv fragment are stabilized by disulfide bridges. In the scFv fragment a short peptide spacer (usually 15–20 amino acids) is introduced in order to link the VH and VL domains covalently. This also allows the possibility of the linkage of two scFv fragments to create "diabodies" which are bispecific in so far as they have two independent antigen binding sites. Bispecific antibodies can also be produced from the fusion of two hybridomas to generate a "quadroma." However, all combinations of light and heavy chains are synthesized in these cells with only a few of the molecules being bispecific. Purification of the required molecules would be a difficult task.

The potential advantage of these recombinant fragments for human therapy is their small size that facilitates tissue penetration, biodistribution, and blood clearance. The fragments can be isolated from libraries of antibodies displayed on the surface of filamentous bacteriophages. This phage display technology is an alternative strategy that can be used instead of mammalian hybridoma technology. The disadvantage is that the recombinant antibody fragments lack glycosylation and also the binding sites for complement and Fc receptors. However, the possibilities exist of conjugating other polypeptide sequences to express the desired effector functions. Conjugation of toxins or specific growth factors to these fragments also allows the future development of immuno-constructs with considerable potential for therapeutic activity.

XVI. THERAPEUTIC ANTIBODIES

Interest in the use of monoclonal antibodies as therapeutic agents has existed for a considerable time. Various monoclonal antibodies have been targeted to membrane-bound proteins specifically expressed in tumor cells. These antibodies can be designed in configurations likely to

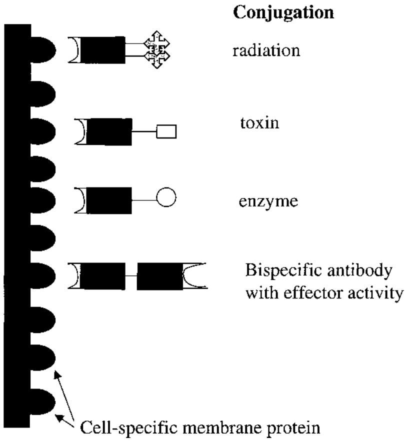


FIGURE 12 Use of antibodies in cancer therapy.

cause the destruction of the target tumor cells (Fig. 12). The conjugation of radioactive or toxic compounds to the antibody can result in a localized high concentration resulting in cytotoxicity. Alternatively, an enzyme may be conjugated that will catalyze the release of a toxic product from an ingested "pro-drug." Another strategy is to promote an effector function through the use of a bispecific antibody that could potentially activate T cells leading to the specific destruction of the target cells.

The rationale behind these methods is to cause localized cell destruction but limit systemic toxicity. However, the results of the initial clinical trials for therapeutic murine antibodies using these strategies was disappointing. This was as a result of unexpected toxicity associated with treatment of immunoconjugates and the undesirable human anti-mouse antibody (HAMA) immune response. However, the administration of antibodies relying on a response within a short period of time period (up to 10 days) were more successful. This included antibodies for radioimaging, radioimmunotherapy or for acute allograft rejection (the OKT3 antibody).

The development of human chimeric antibodies in the 1990s increased rapidly the rate of licensing of monoclonal antibodies as therapeutic agents. These humanized antibodies do not elicit the HAMA effect and also the half-life is much longer than mouse antibodies. Most mouse IgG have a half-life of less than 20 h, whereas an antibody with a human-type constant region can have a half life of up to 21 days. Table III shows eight monoclonal antibodies that have been approved by the U.S. regulatory agency, the FDA for human therapeutic use. As well as these, there are many more antibodies in clinical trials with the expectation that the numbers in therapeutic use will increase in the future.

The success of these chimeric antibodies can be illustrated by Rituxan which is used in the treatment of non-Hodgkin's lymphoma. This has a murine variable region which binds specifically to CD20 on B cells and a human Fc domain to trigger effector mechanisms. CD20 is a protein expressed by over 90% of the lymphoma cells. These tumor cells can become coated by the anti-CD20. This results in activation of the complement pathway and Fc receptor-bearing cells which can destroy the tumor cell.

XVII. ANTIBODIES FROM PLANTS

Antibodies were first expressed in transgenic plants in 1989. Since then various antibody fragments and domains have been produced in plant hosts as well as full-length and multimeric antibodies. The most popular host species for this work has been the tobacco plant, *Nicotiana*, although corn and soybeans have also been utilized. There is no apparent reason why other plants could not be used. The value of using plants for monoclonal antibody production include the absence of animal pathogens, the ease of genetic manipulation, the ability of post-translational modification, and the potential for scale-up to an economic production process.

Transformation involves the stable integration of the appropriate DNA into the plant cell genome. The resulting transgenic plants can be cross-fertilized so as to integrate

TABLE III Monoclonal Antibodies Approved by the FDA for Clinical Use

Antibody	Type	Therapeutic treatment	Company	Date approved
Orthoclone (OKT3)	Murine Ig2a	Allograft rejection	Ortho Biotech	1986
ReoPro	Chimeric (Fab)	Coronary angioplasty	Centocor/ Lilly	1994
Zenapax	Humanized IgG1	Allograft rejection	Protein Design/Hoffman-La Roche	1997
Rituxan	Chimeric IgG1	Non-Hodgkin's lymphoma	Genentech	1997
Synagis	Humanized IgG1	Respiratory synctial virus	Medimmune	1998
Herceptin	Humanized IgG1	Breast cancer	Genentech	1998
Simulect	Chimeric IgG1	Allograft rejection	Novartis Pharm	1998
Infliximab	Chimeric IgG1	Rheumatoid arthritis/Crohn's disease	Centocor	1998–1999

the genes of a multimeric protein structure in a single genetically stable hybrid plant. In one example of the use of this technique, the assembly of secretory IgA was achieved. This is a multimeric immunoglobulin which consists of two Ig units dimerized by a small polypeptide chain (J). Four transgenic plants were produced initially for the expression of a light chain (kappa), a heavy chain, a J chain, and a secretory component. A series of sexual crosses between these plants enabled the generation of a hybrid in which all four components were expressed simultaneously. In this hybrid the four recombinant proteins were assembled into the fully functional secretory immunoglobulin. The antibody expressed in these plants is secreted and accumulates in the apoplasm which is a large aqueous space external to the cells. This is a stable environment unlikely to cause any proteolysis of the accumulated protein.

An alternative method to the production of transgenic plants is the infection of a wild-type plant with a suitable recombinant virus vector. Using this method a monoclonal antibody can be expressed in the leaves of a tobacco plant (*Nicotiana benthamiana*) by infection with two viral vector constructs of tobacco mosaic virus (TMV). In one example, these two constructs contained the genes for the heavy and light chains of an antibody (CO17-1A) against a colorectal cancer associated antigen. A functional full-length antibody was detected in extracts of the leaves of the plants infected with these recombinant viruses. The use of plant virus vectors may have several advantages over the development of transgenic plants. The long generation time associated with plant transformation is avoided. The process also avoids the time-consuming process of crossing transgenic plants to produce hybrids for the expression of proteins with multiple subunits. Different host plants can be infected by the same virus vectors, thus allowing screening for the maximum efficiency of expression.

One major advantage of monoclonal antibodies from plants is the potential low cost of large-scale production. There are commercial companies (such as EPIcyte Pharmaceutical Inc) who are planning clinical trials for plant-produced secretory antibodies for human therapy. These so-called "plantibodies" can be produced at an estimated cost of \$0.01 to \$0.1/mg as opposed to \$1 to \$5/mg for production from cell culture processing of animal-derived hybridomas. The cost of microbial fermentation is lower than that of mammalian cell culture but bacteria lack the ability for efficient multimeric protein assembly and of any post-translational modification. A further potential advantage of the plantibodies is delivery by consumption of plant tissue and thus avoiding any need of purification. These possibilities are particularly applicable in certain cases such as the previously shown ability of a plant-produced antibody against *Streptococcus mutans* to prevent binding of the bacteria to the surface of teeth and thus reducing tooth decay.

Plant cells are eukaryotes and therefore capable of post-translational modification of proteins including N-linked glycosylation. However, although the plant glycan structures have not been analyzed in detail it is likely that these structures are significantly different from those in mammalian systems. For example, the commonly found mammalian terminal sialic acid (N-acetyl neuraminic acid) residue is a structure not found in plants. Also the alpha-1,3 core fucose structure appears to be unique to plants and has been implicated in human allergies to pollen. The potential presence of such unusual glycan structures in plant-derived antibodies might not have an effect on antigen binding but for a therapeutic antibody they are likely to increase the chance of an adverse immunogenic reaction during human treatment. This could limit the use of plant-derived antibodies in certain applications particularly if systemic long-term administration is required.

XVIII. HUMANIZED ANTIBODIES FROM TRANSGENIC MICE

Transgenic mice strains have been produced capable of synthesizing human monoclonal antibodies. Xenomouse strains have large portions of human variable region genes incorporated into the germ line via a yeast artificial chromosome (YAC). The megabase-sized YAC allows the genes for human heavy and light chain immunoglobulin to be incorporated as transgenes into a mouse strain deficient in the production of murine Ig. The large human variable region repertoire incorporated as transgenes allows the mice to generate a diverse immune response comparable to that in humans. These human genes are also compatible with the mouse enzymes that allow class switching from IgM to IgG. The immunoglobulin generation will also undergo somatic hypermutation and affinity maturation, a natural process that enhances the affinity of the antibody for the antigen. Thus an antigen introduced into a Xenomouse produces a human monoclonal antibody with high specificity for its corresponding antigen. Such antibodies have already proved their potential in clinical trials.

XIX. THE IMPORTANCE OF GLYCOSYLATION TO THERAPEUTIC ANTIBODIES

The glycosylation pattern of the immunoglobulin structure has particular relevance to the production and use of

monoclonal antibodies as therapeutic agents. Any *in vitro* production process results in a heterogeneity of glycan structures of the product protein. To avoid any undesirable immune response in the use of such antibodies it is important to maximize the content of fully processed glycans. There are various parameters that affect the glycan processing from the metabolic profile of the hybridoma to the environmental conditions of culture. It has been shown that the glycan structures vary with the specific activity of key glycosylating enzymes contained in a hybridoma. This in turn depends upon the enzymic profile of the parental cell lines used in the hybridization process. In one study hybridomas were produced from parental cell lines only one of which had an enzyme (UDP-N-acetylglucosamine: β-D-mannoside β-1,4-N-acetylglucosaminyltransferase; GnT-III) responsible for the addition of a bisecting GlcNac. As expected the resulting hybridomas had varying levels of GnT-III. Of interest was the fact that the content of bisecting GlcNAc in the antibodies produced by each hybridoma was a reflection of the intracellular activity of GnT-III.

A further example of the effect of the producer cell line on the glycosylation pattern of a monoclonal antibody has been shown for the IgG, CAMPATH-1H. This is a humanized recombinant murine monoclonal antibody developed for human therapy and has been expressed in various cell lines including a murine myeloma and Chinese hamster ovary (CHO). The glycosylation of the antibody produced from CHO was consistent with normal human IgG. That is fucosylated biantennary structures containing zero, one, or two galactose residues. However, the immunoglobulin from a murine myeloma (NS0) results in some potentially immunogenic glycoforms containing Galα(1-3)Gal terminal residues. Such hypergalactosylated proteins have been shown following expression from various murine cells.

Cell culture conditions have also been shown to affect product glycosylation. Relevant culture parameters include the accumulation of ammonia, the dissolved oxygen level, glucose depletion, lipid composition, pH and protein content of the medium. The glycosylation of monoclonal antibodies from hybridomas is particularly susceptible to the dissolved oxygen level of the culture, which on a large scale is often maintained at a specific set-point through an oxygen probe that controls the gaseous input to the bioreactor. This dissolved oxygen (DO) is usually calibrated from 0 to 100%, which is the level of oxygen relative to saturation with air. The three predominant glycoforms found in immunoglobulin are shown in Fig. 13. The relative proportion of these glycoforms has been found to change depending upon culture conditions. For example, the dissolved oxygen level of a hybridoma culture has a noticeable effect on this distribution. Whereas a normal distribution of G0, G1, and G2 is found at 50% DO, lower levels of oxygen (<10%) may lead to poor galactosylation and a consequent increased proportion of G0. The data shown in Fig. 14 show a proportional decrease in G0 and an increase in G2 at higher DO levels.

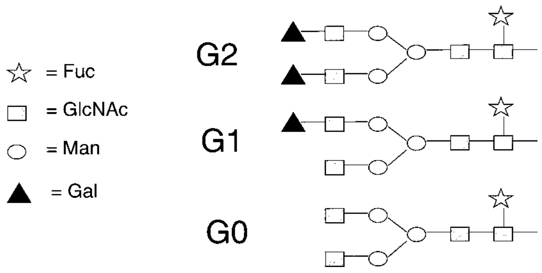


FIGURE 13 Glycan structures of IgG.

XX. LARGE-SCALE PRODUCTION OF MONOCLONAL ANTIBODIES FROM HYBRIDOMAS

A. *In Vivo* (Ascites) Production of Monoclonal Antibodies (Mabs)

One of the original methods employed for the large-scale production of monoclonal antibodies was to grow the selected hybridoma cell lines *in vivo*, following injection into the peritoneal cavity of mice. The hybridomas grow essentially as tumors in a liquid milieu termed the ascites fluid. The secreted antibodies are then extracted from the aliquots of this fluid. After 5–21 days, the peritoneum is tapped for the antibody-rich fluid. This has been a standard method for the production of the thousands of monoclonal antibodies that are now available commercially. The antibodies may be required in variable quantities from milligrams to kilograms. For laboratory reagents the

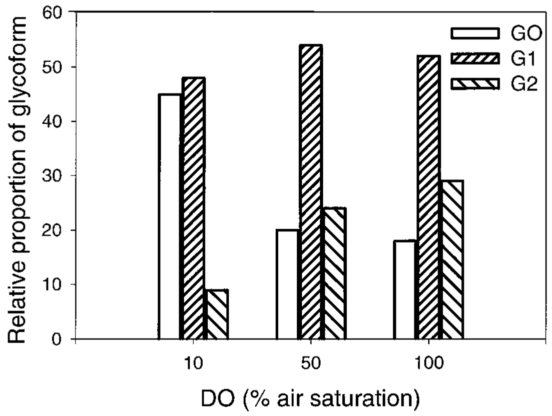


FIGURE 14 Effect of dissolved oxygen on the glycoform profile of IgG.

typical minimum quantity for sale is 200 μg which may cost around $300. This method of production is convenient because variable quantities of different antibodies can be produced from batteries of mice.

The production of monoclonal antibodies *in vivo*, using mice (or other laboratory animals) has come under increasing criticism because of the ethical issues posed by the use of laboratory animals. In Europe, regulatory approval of this method has been withdrawn except for cases were alternative methods are shown not to be available. Mice generating Mabs typically exhibit abdominal distention, anorexia, anemia, decreased activity, and body mass, dehydration, difficulty in walking, respiratory distress, shock, hunched posture, peritonitis, immunosuppression, and possibly death.

Additionally, the ascites method presents problems for product purification. The overall protein content of ascites fluid is high, posing considerable difficulty in obtaining a pure monoclonal antibody. Furthermore, the ascites fluid contains antibodies secreted by the host mouse and these are virtually impossible to separate from the monoclonal antibody. Thus the final "purified" product has residual activity that may interfere with the application of the monoclonal antibody.

B. *In Vitro* Production

The basis for commercial production of monoclonal antibodies from hybridomas is cell culture technology which involves the growth of isolated mammalian cells in liquid culture *in vitro*. Cells are grown in bioreactors, and can be produced in high densities if the appropriate physical conditions and nutrients are provided. Small volume cultures (<200 ml) are usually set up in T flasks or spinner flasks in temperature-controlled incubators, often without controlling other culture parameters. However, in order to obtain high cell densities and maximize productivity of secreted products at a larger scale (>1 liter) other culture parameters are controlled. These include oxygen supply, temperature, pH, and culture mixing. Three culture system bioreactor designs have been used routinely for the production of monoclonal antibodies. These include stirred-tank, air-lift, and hollow fiber bioreactors.

C. Stirred-Tank Bioreactor

The stirred tank bioreactor is a simple and widely used fermenter design that consists of a cylindrical vessel with a stirrer. The design has been used extensively in all microbial fermentation and has been the main system used in yeast fermentation in the brewing industry for centuries. However, animal cells are more fragile and grow more slowly than most bacteria or fungi. They require gentler culture conditions and control systems that are optimized for lower metabolic rates. Therefore, the design, mode of operation, and control systems of a stirred tank reactor used for animal cells are distinctly different from those that would be applicable to bacterial or fungal cells.

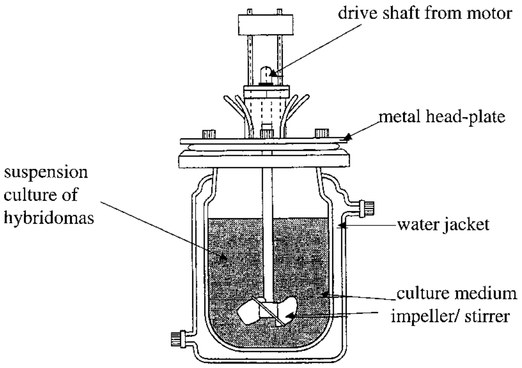


FIGURE 15 Stirred-tank bioreactor.

The stirred tank reactor has been developed commercially in large-scale animal cell culture processes up to a volume of at least 10,000 liters. For laboratory use there are also numerous bench-top stirred tank reactors (1–5 liters) that are available commercially and that have been designed specifically for the growth of animal cells in suspension. Figure 15 shows a typical design. Bench-top models are generally made of glass with a stainless-steel head-plate, whereas the larger fermenters are made entirely from stainless steel. The metal head plate of a stirred tank reactor consists of a range of ports and pipes. This allows electrodes to be inserted and tubing to be attached for media input or sampling.

Manufacturers of bench-top models include the following: Applikon, New Brunswick, LH Fermentation, Setric SGI, Braun, Bio-engineering. Each of these companies produces uniquely designed and controlled fermenters, all of which have been shown to be suitable for animal cell culture.

D. The Airlift Fermenter

This type of fermenter consists of a tall column with an inner draught tube (Fig. 16). Fluid circulation is provided by a stream of air which passes through the inside of the draught tube. This is a simple system without mechanical components and therefore not susceptible to breakdown. Bubble or foam damage is minimized by having a long column, since it has been shown that maximum cell damage occurs at the point of bubble bursting at the top of the liquid column. Airlift fermenters (>1,000 liters) have been used routinely for the production of bulk quantities of monoclonal antibodies from hybridomas.

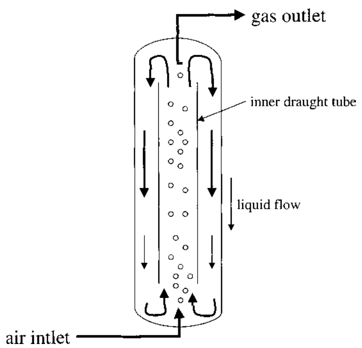


FIGURE 16 Airlift fermenter.

E. The Hollow Fiber Bioreactor

This consists of bundles of synthetic, semipermeable hollow fibers which offer a matrix for cell growth similar to the vascular system *in vivo*. Liquid can flow through the fibers (the intracapillary space) or through the space between the fibers (the extracapillary space). In the normal operation culture medium is pumped through the intracapillary space and a hydrostatic pressure permits the exchange of nutrients and waste products across the capillary wall. The cells and large molecular weight products are held in the extracapillary space (Fig. 17).

A major limitation of this type of system is that the pressure difference that may establish along the length of fibers can cause nutrient gradients and uneven cell growth. Such pressure differences and gradients become an increasing problem with scale-up. The design of some hollow fiber systems is intended to correct this problem. Here the pressure differential between the intra- and extracapillary space is continuously monitored by sensors which serve to control the opening and closing of valves which in turn affect the capillary pressure.

The exchange of molecules through the fiber wall occurs in phases governed by a cyclic mode of pressure changes which prevent undue gradients developing along the fibers. The hollow fiber system is suitable for both anchorage-dependent and independent cells. Continuous operation allows a high rate of product recovery from a stationary high-density culture held in the extracapillary space over a long period of time.

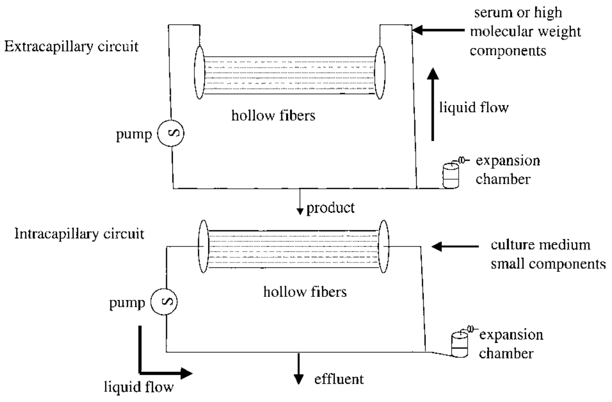


FIGURE 17 Hollow fiber bioreactor.

XXI. THE CONTROL OF CULTURE PARAMETERS

There are several culture parameters that are important to control for maximum cell growth and antibody production. These include agitation, temperature control, pH control, and oxygen supply.

A. Agitation

Animal cells tend to be fragile compared to fungal or bacterial cells. Cells in suspension can be damaged by various forces acting in a stirred culture, the major damaging force is from bubble bursting on the culture surface resulting from culture aeration. The hydrodynamic shear force resulting from the motion of a stirrer is thought to be of lesser importance; nevertheless the stirring speeds commonly adopted for animal cell cultures are considerably lower than those for bacterial cultures of equivalent volume.

The simplest stirring operation involves the rotation of a suspended bar by a magnetic stirrer. This is the system used in glass spinner bottles and is suitable for stirring cultures up to a volume of 1 liter. At larger volumes, such magnetic stirrers are not suitable because of the increased energy required for rotation. Top-drive mechanical motors are normally used for stirred tank reactors from bench-top models to the larger commercial fermenters. In the design shown in Fig. 15 the stirring shaft fits through the stainless-steel head-plate and into the sterile culture. The stirring motor and outer part of the drive shaft can normally be disconnected from the head-plate to allow the fermenter to be autoclaved. In early fermenter models, the stirring shaft was connected through the head-plate by replaceable rubber/silicone seals, which were vulnerable to damage and provided entry points of contamination. Later models have sealed units which are far more reliable.

Typically, maximum stirring rates of 100–150 rpm are used for cells in suspension. In order to ensure adequate mixing at low stirring speeds, the culture vessels are designed with a round bottom, which distinguishes them from the flat-bottomed bacterial fermenters. Impeller blades which are fitted at the end of mechanical drives shafts are designed to allow vertical as well as horizontal

liquid flow. The pitched-blade or marine type impellers are particularly suitable.

B. Culture pH

The optimal pH for hybridoma growth is around pH 7.0–7.4 and maximum cell growth occurs if this is maintained. A pH probe can be used to control the sparging of cultures with CO_2 or by low volume additions of a concentrated sodium bicarbonate solution. Bicarbonate is used in preference to sodium hydroxide because of the danger of over-shooting the set-point with the stronger alkali. These additions are normally governed by a computer-controlled pump or gas valve to a preset pH value. Acid (e.g., HCl) may also be added from an external source. However, this is usually not required as lactic acid is normally produced by cell metabolism and causes the culture pH to decrease during growth. The culture pH in a bioreactor is often controlled by the rate of CO_2 gas flow into the culture. An electronic pH controller regulates the on/off function of a control valve.

C. Oxygen

The supply of oxygen to satisfy cell metabolism is one of the major problems associated with fermenter scale-up. The oxygen consumption rate of hybridoma cells varies from 0.06 to 0.6 mmol/liter/hr for a culture at 10^6 cells/ml. In small cultures (T-flasks, etc.), the oxygen demand can be satisfied by gas diffusion from the head space through the culture surface. However, with increasing culture volume, the surface to volume ratio decreases. At cultures of 1 L and above, the surface/volume ratio is generally too low to satisfy the overall oxygen demand at this cell concentration.

Because the solubility of oxygen is low and a continuous supply of oxygen is required to satisfy the cellular metabolism at higher culture volumes,. The maximum concentration of oxygen in culture media which is in equilibrium with air is 0.22 mM at 37°C—referred to as "100% air saturation." Growth of many animal cell lines has been found to be optimal at dissolved oxygen levels (DO) below the maximum oxygen solubility and corresponding to 20–50% of air saturation. This DO level may be maintained by a control system that incorporates an oxygen probe and input of sparged air or oxygen.

An added problem is that excessive gas sparging may cause cell damage particularly in cultures with a large surface to volume ratio. This problem may be offset by the use of chemical protectants such as the polymer, Pluronic F-68, or by alternative methods of introducing oxygen into the culture. The alternative methods may include gas sparging in a media reservoir not in contact with the cells or via a the surface of silicone tubing introduced into the culture vessel.

XXII. SERUM AND SERUM-FREE MEDIUM FOR ANTIBODY PRODUCTION FROM HYBRIDOMAS

The growth of mammalian cell lines requires a chemically defined liquid basal medium which is traditionally supplemented with 10% (by volume) bovine serum to provide supplements of growth factors. Although bovine (or other animal) serum provides excellent growth support for cells in culture, there are significant disadvantages in using serum as a an additive to hybridoma cultures. These include:

- Batch-to-batch variation in composition. To generate a supply of bovine serum for use in cell culture, the serum is extracted from a herd of cows and pooled. Each batch is then identified by the supplier and can be sampled by the buyer for suitability in an industrial process. However, the composition of each batch varies in undefined ways depending upon the diet of the cows. This variation can cause significant differences in the growth-promoting characteristics of the serum, and ultimately causes significant differences in productivity of the cell culture process.
- A high protein content that hinders product purification. The cells grown in a bioreactor secrete the product of interest (normally a protein) into the culture medium. If the culture medium contains serum, its protein concentration is already high. Serum has a protein concentration of 60–80 mg/ml and so the basal level of protein in a 10% serum-supplemented culture medium is 6–8 mg/ml. In comparison, the concentration of a protein secreted by the cells being cultured typically reaches 100–150 μg/ml and therefore a difficult purification process is required to separate the product from the serum protein (called downstream processing). The monoclonal antibody of interest may well be mixed with any other antibodies present in the serum and these are virtually impossible to separate. The disadvantage of an impure or poorly purified product such as a monoclonal antibody is that there may be unwanted side reactions connected with its use as a diagnostic or therapeutic agent that reduces its effectiveness.
- The potential for product contamination. The threat of contamination arises from unwanted viruses and mycoplasma that may be present in serum as well as the undefined and uncharacterized prion agents of bovine spongiform encephalopathy (BSE, or "*Mad Cow Disease*"). Because of the concern over the

potential human consequences of the presence of these contaminants in therapeutic products, most regulatory authorities have demanded the use of serum-free processes for the production of therapeutic products when available.

A. Serum-Free Media

Given the disadvantages described above, the growth of cells in bioreactors using *serum-free media* offers an alternative solution:

- Serum-free media reduce the risk of exposure to agents of zoonotic diseases, like BSE, by being devoid of animal-derived components.
- Serum-free media can be formulated with a low protein content to offer enhanced purity and higher quality of the final cultured product.

Historically, serum-free media have exhibited poor growth characteristics compared to serum-supplemented media, and for this reason have not been widely used as a replacement to serum-supplemented media. However, serum-free media technology is continuously improving and serum-free media products have been shown to exhibit growth and productivity characteristics that are comparable or superior to serum-supplemented media.

XXIII. CONCLUSIONS

The ability to produce monoclonal antibodies from hybridomas emerged from a technology developed in the early 1970s and reported in 1975. Since then monoclonal antibodies have found wide application in research and in diagnostic tests because of their high specificity in recognizing antigens. However, the therapeutic application of monoclonal antibodies has taken a long time because of a range of side-effects associated with undesirable immune responses in humans of murine-derived antibodies. The situation is now rapidly changing with the ability to produce humanized or fully human antibodies. This has enabled the approval of monoclonal antibodies for a range of therapies including transplantation, cancer, infectious disease, cardiovascular disease, and inflammation. There are presently eight antibodies approved by the FDA for therapeutic use (see Table III) with several hundred awaiting the results of clinical trial. Because these approved antibodies are human (or humanized) immunoglobulins they enable effector functions to direct complement-dependent cytotoxicity to a target cell. Other biological effects are also possible by conjugation of compounds to the antibody.

Therapeutic antibodies are required in much larger quantities than those used in diagnosis or as laboratory reagents. Therefore, it is certain that the requirements for large-scale production of hybridomas will increase. There is a need to ensure that the conditions of culture are compatible with full and appropriate human glycosylation profiles of the synthesized immunoglobulins. Therefore, work to understand fully those conditions that allow this to take place will continue.

SEE ALSO THE FOLLOWING ARTICLES

Gene Expression, Regulation of • Immunology—Autoimmunity • Mammalian Cell Culture • Metabolic Engineering • Nucleic Acid Synthesis • Protein Folding • Protein Structure • Protein Synthesis • Tissue Engineering • Translation of RNA to Protein

BIBLIOGRAPHY

Borrebaeck, C. A. K., and Hagen, I. (eds.) (1993). “Electromanipulation in Hybridoma Technology: A Laboratory Manual,” Stockton Press & W. H. Freeman/OUP, New York.

Butler, M. (1996). “BASICS: Mammalian Cell Culture and Technology,” Oxford University Press, Oxford.

Cambrosio, A., and Keating, P. (1996). “Exquisite Specificity: The Monoclonal Antibody Revolution,” Oxford University Press, Oxford.

Delves, P. J. (ed.) (1994). “Cellular Immunology Labfax,” Academic Press, London.

Malik, V. S., and Lillehoj, E. P. (eds.) (1994). “Antibody Techniques,” Academic Press, London.

Mather, J., and Barnes, D. (eds.) (1998). “Animal Cell Culture Methods,” Academic Press, London.

Mizrahi, A. (ed.) (1989). Adv. in Biotechnological Processes Vol. 11, “Monoclonal Antibodies: Production and Application,” A. R. Liss, New York.

Seaver, S. S. (ed.) (1986). “Commercial Production of Monoclonal Antibodies,” Marcel Dekker, New York.

Springer, T. A. (ed.) (1985). “Hybridoma Technology in the Bioscience and Medicine,” Plenum Press, New York.

Wang, H. Y., and Imanaka (eds.) (1999). “Antibody Expression and Engineering,” Oxford University Press, Oxford.

Harbour, C. and Fletcher, A. (1991). “Hybridomas: production and selection,” *In* “Mammalian Cell Biotechnology: A Practical Approach” (M. Butler ed.), pp. 109–138, Oxford University Press, Oxford.

James, K. (1990). “Therapeutic monoclonal antibodies—their production and application,” *In* “Animal Cell Biotechnology,” (R. E. Spier and J. B. Griffiths, eds.), Vol. 4, p. 205, Academic Press, London.

McCullough, K., and Spier, R. E. (1990). “Monoclonal antibodies in biotechnology: Theoretical and practical aspects,” *In* “Cambridge Studies in Biotechnology,” Vol. 8, Cambridge University Press, Cambridge.

Fukuta, K., Abe, R., Yokomatsu, T., Kono, N., Nagatomi, Y., Asanagi, M., Shimazaki, Y., and Makino, T. (2000). “Comparative study of the N-glycans of human monoclonal immunoglobulins M produced by hybridoma and parental cells,” *Arch. Biochem. Biophys.* **378,** 142–150.

Hiatt, A., Cafferkey, R., and Bowdish, K. (1989). “Production of antibodies in transgenic plants.”

Image-Guided Surgery

Ferenc A. Jolesz
Harvard Medical School

I. Image-Guidance Methods and Technologies
II. Intraoperative Imaging
III. Intraoperative Magnetic Resonance Imaging
IV. Image-Guided Neurosurgery
V. Thermal Ablations
VI. Cryoablation
VII. Conclusion

GLOSSARY

Cryosurgery Surgery that involves the application of extremely low temperatures to destroy tissue.

Focused ultrasound surgery (FUS) Surgery that involves the use of extremely high frequency sound targeted to highly specific sites of a few millimeters or less.

Image-guided therapy (IGT) A general term for therapies based on some form of imaging data describing conditions below the visible surface, as a complement to or replacement for direct observation.

Interstitial laser therapy (ILT) A form of thermal ablation using a laser as the heat source.

Intraoperative imaging Imaging done during the actual course of an operation.

Magnetic resonance imaging (MRI) A contemporary technique of medical imaging in which a rapid oscillation of atomic nuclei occurs when certain elements placed in a strong magnetic field are exposed to a radio pulse of appropriate frequency; this produces a signal that can be detected by external sensors and employed to provide images of subsurface tissue.

Minimally invasive surgery Surgery that involves the least possible intrusion into the body of instruments or other foreign material and mechanisms.

Stereotactic surgery Surgery that involves the highly specific location of a discrete site (e.g., in the brain) and the precise direction of the surgical device to and at that site.

Thermal ablation The removal or destruction of tissue by means of heat.

CURRENT EXCITING progress in minimally invasive surgery, the introduction of new imaging modalities, and the availability of high-performance computing are the setting for the development of novel *image-guided therapies*. Image-guided therapy (IGT) is changing the fundamentals of traditional surgery by replacing and/or complementing direct visualization with volumetric imaging (Jolesz, 1997). This new approach not only represents a technical challenge but also a transformation of procedures based on hand–eye coordination into interactive

navigational operations. The presentation of multimodality-based images has to be merged into a single model in which anatomy and pathology have to be distinguished but integrated into the same intuitive framework. Therapy systems have to be linked with imaging systems to constitute a complete therapy delivery system. At the same time a multidisciplinary team has to be created which combines surgeons, interventionalists, imaging experts, and computer scientists.

Such an environment is radically different from the conventional operating room. The surgeon's view of the surface of the operational field is coupled by images showing what is beyond the visible surface. This leads to a definite change in surgical approaches and methods and results in a close integration of image-based information with surgical procedures. This new integrated setting is not optimized yet and it has been the subject of intense research (Jolesz and Shtern, 1992). The overall goal of IGT is to integrate all the accessible information (both preoperative and intraoperative imaging data) into a single complete operational *therapy delivery system*.

These therapy delivery systems can be suited to different applications. These call for interdisciplinary teams from various clinical specialties to work together with engineers and computer scientists using not only surgical and radiological methods, but biomedical engineering principles in the process of combining imaging and therapy devices. It is anticipated that this emerging field will embrace 1ess-invasive therapy options and will result in better clinical outcomes and reduced cost.

I. IMAGE-GUIDANCE METHODS AND TECHNOLOGIES

IGT systems use preoperatively acquired images to create models which can be used for localization, targeting, and visualization of the three-dimensional (3D) anatomy. These volumetric models support *surgical planning* and/or simulations to define and optimize access strategies and to simulate planned trajectories (Kikinis *et al.*, 1996). When these models are registered to the patient's actual anatomy, they relate image-based coordinate locations in the surgical field. The use of sensors or tracking devices enables the surgeon to navigate and carry out procedures using all the preoperative multimodality-based information, which is tied together in the model. Acquisition, display, and visualization of an image in surgery is different from that in diagnosis. Preoperative images can be acquired with optimal quality and without seriously limiting the imaging time. Intraoperative guidance, however, limits the time acceptable for imaging. To fully integrate acquisition and display, intraoperative imaging has to be dynamic and primarily defined by the procedure (Fig. 1).

The navigational systems display images interactively with orientation and location defined by the position of sensors attached to surgical instruments or other tracked devices. Interactive image guidance can direct the surgeon to the target using on-line trajectory optimization. By displaying alternative access routes, surgical planning can be performed at the operating room table. Navigational systems (frameless stereotaxy) have a relatively good accuracy and they are feasible unless the surrounding anatomy is changing significantly during the procedure.

In combination with navigational and interactive display tools preoperative models can aid a variety of diagnostic and therapy applications. For diagnosis, the primary objective is detection and description of a lesion. For therapy, the purpose is localization and targeting. For surgery and targeted therapy, it is essential to know the exact position and 3D extent of the abnormality and its relationship to the surrounding anatomy. The representation of anatomic structures and their functions along the

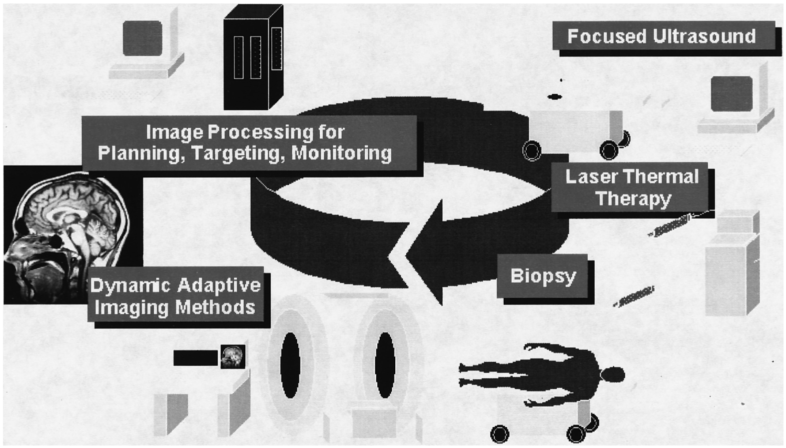


FIGURE 1 The concept of image-guided surgery.

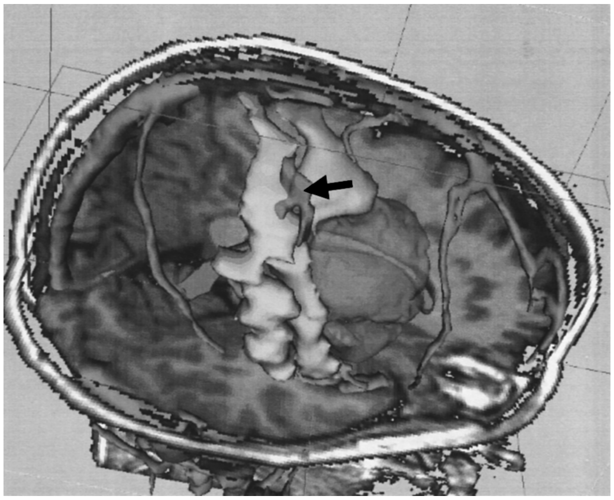


FIGURE 2 Image fusion. Two-dimensional MR images of the brain are combined with three-dimensional information obtained from functional MR and MR angiograms. The red represents intracranial blood vessels. The green corresponds to a brain tumor. The purple represents the pre-central gyrus (motor cortex) and the yellow corresponds to the post-central gyrus (sensory cortex). The arrow points to the area activated by finger tapping recorded by functional MRI.

path or trajectory of the surgical device is important for targeting. Image-guidance tools should provide 3D representations of both the target and the entire operational volume (Fig. 2).

There are several unresolved basic biomedical engineering questions in IGT. Most of the efforts so far have been concentrated on image processing methods including various registration and segmentation approaches. Most of the applications of IGT require robust algorithms and automated methods that create patient-specific models of relevant anatomy from multimodal imaging. The process of selecting tissue components with anatomic or pathologic importance is called *segmentation* (Cline *et al.*, 1990; Gibson *et al.*, 1998; Held *et al.*, 1996; Wells *et al.*, 1996a). The other important computerized method that aligns multiple datasets with each other and with the patient is called *registration* (Pelizzari *et al.*, 1989; Grimson *et al.*, 1996; Wells *et al.*, 1996b). Both techniques may utilize shape description methods for capturing morphology and its biological variation. The challenge is to integrate these technologies into complete and compatible IGT systems. The ultimate goal is to create the computational infrastructure and an associated suite of methods to support a broad range of procedures (Warfield *et al.*, 1998).

II. INTRAOPERATIVE IMAGING

The main purpose of IGT is the integration of anatomic and functional information with surgical and other therapy methods. The availability of patient registered, continuously updated "fused" multimodal information in an intraoperative setting increases safety and may result in better outcome by reducing the invasiveness of the procedures, decreasing complications, and increasing the effectiveness of surgery. Image-based information can be utilized accurately to target and cut out diseased tissues and at the same time avoid critical structures. During surgery most of the structures and the related functions are unseen by the surgeon but can be displayed interactively.

Intraoperative shifts and deformations are unavoidable and mostly unpredictable. These displacements are the results of mechanical factors, physiologic motions, and pathophysiologic processes like edema or hemorrhage. The unwanted movement of tissues and the reduction or swelling of tissue volumes by the advancing surgery can be so substantial that the use of preoperatively acquired images for guidance become impossible. The only solution to this problem is intraoperative imaging, which updates the original preoperative (baseline) 3D model. The potential use of algorithmic tools which model rigid and nonrigid deformations is limited and only volumetric intraoperative imaging can provide correct, updated information (Cotin *et al.*, 1999; Cover *et al.*, 1993; Hata *et al.*, 1998).

The application of intraoperative image guidance for monitoring and controlling open surgeries, endoscopic procedures, thermal ablations, brachytherapy, and targeted drug delivery can consolidate minimally invasive therapies. IGT methods have already had an impact on the fields of interventional radiology, radiation oncology, and surgery. In the future a strong coordinated multifocused, multidisciplinary translational research effort is necessary to promote the development and implementation of image-guided interventions. This requires innovative approaches, novel applications, and the more efficient use of computer technologies. There is also a need for more advanced therapy devices and for a more complex and diverse technological infrastructure. Examples of current integrated IGT systems and their clinical application are described below.

III. INTRAOPERATIVE MAGNETIC RESONANCE IMAGING

Interactive intraoperative MRI (IMRI) guidance allows one accurately to localize and target in order to optimize surgical approaches that avoid critical structures and decrease the vulnerability of surrounding functionally active normal tissues (Fig. 2). In addition, by measuring specific functional (perfusion, flow) or physical (diffusion, temperature) parameters MRI can monitor and/or control energy delivery, targeted drug delivery, or other therapy methods. Since the introduction of interventional and intraoperative

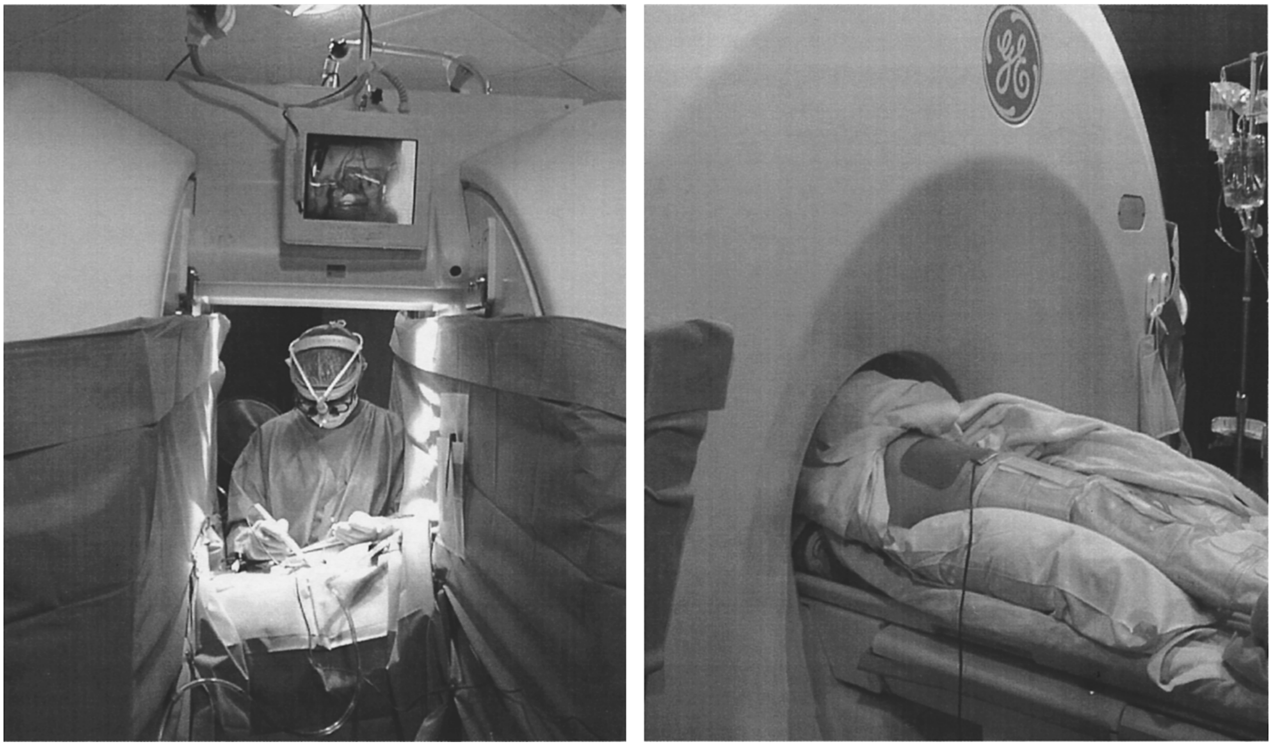


FIGURE 3 Open-configuration intraoperative MR imaging system (Signa SP, GE Medical Systems). Left: the surgeon operates between the two components of the magnets. There is full access to the patient. The surgical instruments are tracked using optical sensors to guide interactive imaging (display seen at top). Right: the patient is introduced into the magnet.

magnetic resonance imaging (Figs. 3 and 4), considerable experience has been amassed (Jolesz, 1994; Schenk *et al.*, 1995; Jolesz, 1998) (Fig. 4). The IMRI integration has been a good exercise in blending together various components of IGT (Silverman *et al.*, 1998). Several surgical (brain, spine, breast) and interventional clinical applications, such as MRI-guided endoscopy (Fried *et al.*, 1998; Hsu *et al.*, 1998), interstitial laser therapy (Kettenbach *et al.*, 1998; Vogl *et al.*, 1997; Kahn *et al.*, 1998), cryoablation (Silverman *et al.*, n.d.), MRI-guided focused ultrasound treatment (Cline *et al.*, 1994; Chung *et al.*, 1996; Hynynen *et al.*, 1996, 1997, n.d.), and MRI-guided brachytherapy (D'Amico *et al.*, 1998) have been tested.

The integration of combining the operating room with MRI and high-performance computing is necessary. The intraoperative MRI environment and its clinical utilization show the way that IGT can be applied to several procedures. The combination of pre- and intraoperative image acquisition, on-line image processing, and intraoperative display utilizes all the available intraoperative and preoperative information. The skillful integration of the software and hardware components of intraoperative MRI, navigational tools, and multimodality imaging

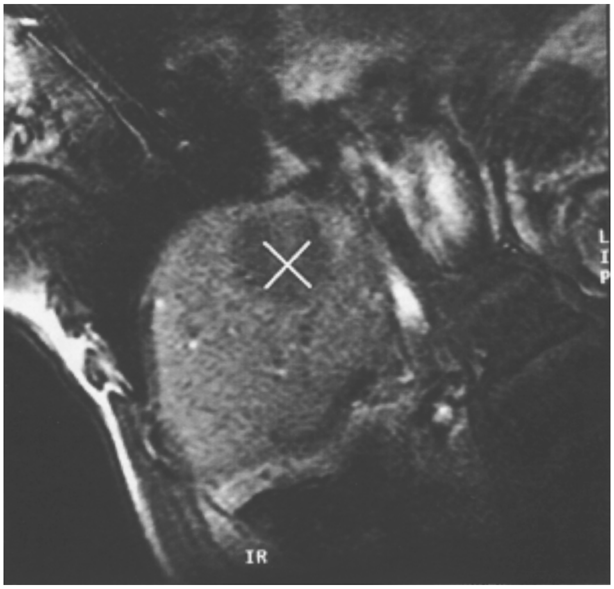


FIGURE 4 Localization of a liver tumor in the interventional MR system, using virtual needle tip (marked by cross-hairs). Targeting is accomplished by tracking the position of the needle holder and displaying image planes defined by the position of the probe.

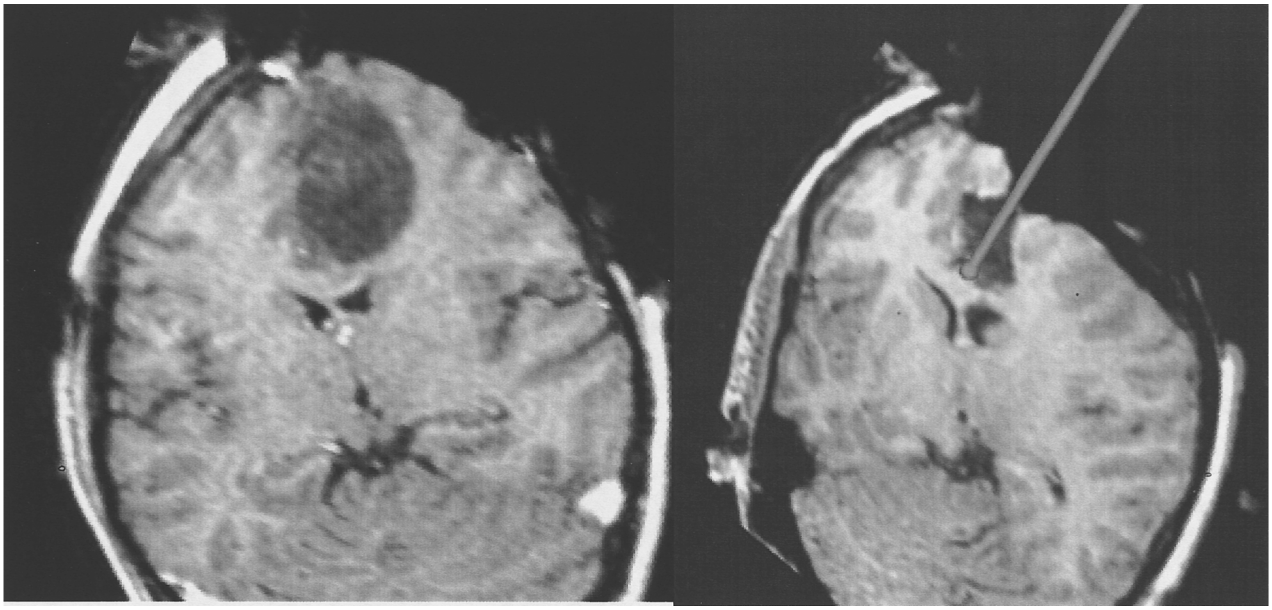


FIGURE 5 Navigational guidance during brain surgery within the intraoperative MRI. Left: the appearance of the brain (with large frontal tumor) immediately after craniotomy. Right: image of the brain after tumor resection. The position of the tracked pointer is seen in the resection cavity. Note the significant deformation of the brain in the advanced stage of surgery.

and intraoperative display guarantees that intraoperative trajectory optimization and navigation can be accomplished in a user-friendly environment (Fig. 5). The integration also includes the engineering setting of high-performance computing and the use of the hospital network.

The IMRI methods require high spatial and temporal resolution, multiplanar imaging, interactive navigation, and a 3D visualization (Figs. 6 and 7). MRI represents excellent tissue characterization for both anatomy and pathology. The multiplanar and volumetric imaging permits the understanding of three-dimensional anatomic relationships. The spatial resolution is appropriate to achieve the accuracy accepted in stereotactic neurosurgery. The temporal resolution of MRI is around the 1-sec range using fast and ultrafast imaging sequences. These fast imaging methods allow close to real-time imaging in the presence of physiological motion, sufficient to track instruments and follow the changes induced by therapy interventions. In addition, specific MRI pulse sequences can

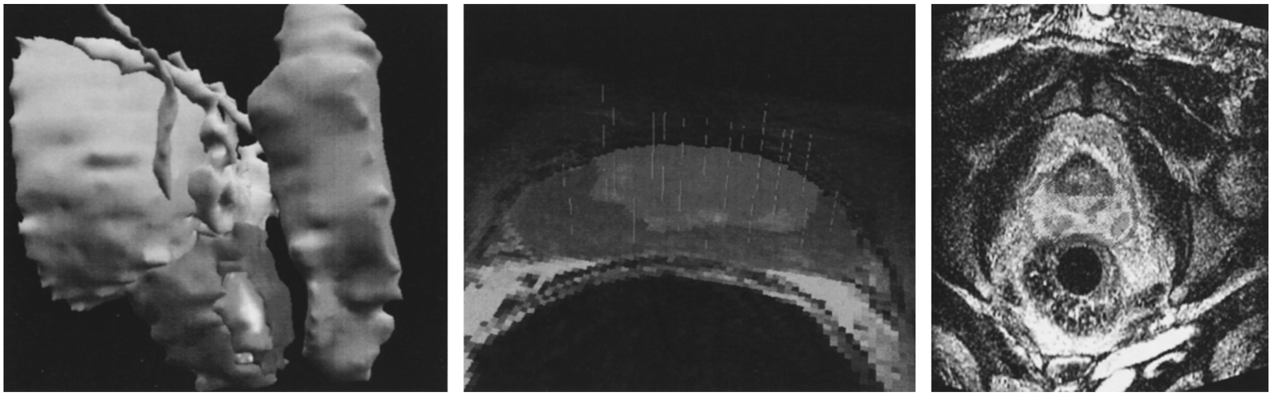


FIGURE 6 Planning and execution of MRI-guided prostate brachytherapy. Left: three-dimensional model of the prostate, tumor, rectum, bladder, and seminal vesicles. The pelvic anatomy was segmented based on MR images. Middle: the planning of the procedure. The peripheral zone and central zone of the prostate are depicted on a cross-sectional MRI slice acquired with an endorectal coil. The dashed lines represent individual needle trajectories. Right: display of dose distribution on two-dimensional MRI of the prostate.

be developed to utilize dynamic imaging methods, such as MR fluoroscopy, keyhole imaging, and adaptive imaging.

Since its initial introduction IMRI has matured from a research tool into a clinical approach that can transform minimally invasive surgery and interventional radiology into a more advanced stage. Currently several open-configuration and short-bore MRI systems suitable for some percutaneous procedures and/or open surgeries are being marketed. In areas like neurosurgery and endoscopic surgery, MRI-based guidance systems may provide more effective treatment options than conventional surgery or other image-guidance techniques such as ultrasound and X-ray computed tomography. Intraoperative MRI may result in improved patient care, reduced invasiveness, and a safer surgical or interventional procedure. MRI-based image guidance may induce the development and implementation of new surgical approaches. This is a challenging new technology which can lead to significant changes in surgical procedures and other treatment methods.

The future of IMRI depends not only on the evolution of MR imaging technologies, but also on the successful integration of computers and therapy devices. Surgeons and radiologists using intraoperative MRI have full access to all available preoperative image based information and on-line MRI can update this baseline information. Interfaces between the operators and imaging and therapy systems are necessary to control the flow of information. Anatomy-, function-, and therapy-induced changes should be displayed in an integrated way.

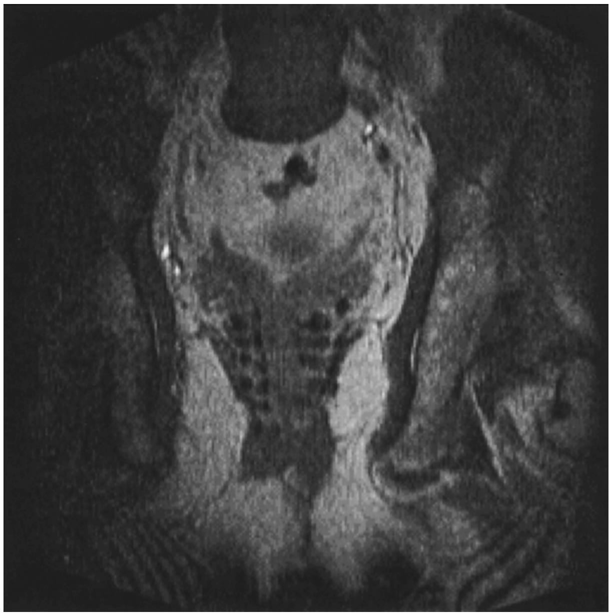


FIGURE 7 The position of the radioactive seeds implanted within the prostate as is seen in coronal MR images.

IV. IMAGE-GUIDED NEUROSURGERY

Over the past decade a distinct field of neurosurgery, image-guided neurosurgery, has evolved through advances in neuroimaging, computer science, and frameless stereotactic techniques. Image-guided neurosurgery introduces effective neuroimaging technologies into the operating room by utilizing advanced computing and engineering technology. Using various intraoperative display tools and interfaces, navigational guidance is applied for localization and targeting (Grimson *et al.*, 1996; Maciunas *et al.*, 1992; Heilbrun *et al.*, 1992; Laborde *et al.*, 1992; Galloway *et al.*, 1992; Barnett *et al.*, 1993; Zamorano *et al.*, 1993; Zinreich *et al.*, 1993). Computerized image-guidance methods and navigational tools have not been tested or carefully evaluated and there has been lack of appropriate methods to assess and validate the complex machinery used; in particular, there has been no attempt to relate this technology development to clinical outcome measures.

A significant limitation of current image-guided neurosurgery is that it is based on preoperative models, which cannot be updated intraoperatively. As the neurosurgical procedure advances, the brain can deform or shift substantially due to the surgical insult and operation. This limits the usefulness of the baseline preoperative information. Intraoperative imaging update can compensate for the changes. Navigational tools built into the intraoperative imaging systems (ultrasound, CT, MRI) permit interactive imaging guidance for biopsies or surgery (Nakajima *et al.*, 1997; Bucholz *et al.*, 1993). Although the image quality of ultrasound has been improving, it is still of lesser value in comparison with MRI. Nevertheless further advances in technology may change the current situation (Koivukangas *et al.*, 1993). The use of CT is limited by ionizing radiation and tissue differentiation. It is inferior to MRI because lacks multiplanar imaging capabilities, high contrast, spatial resolution, and high sensitivity.

With intraoperative MRI guidance one can identify surgical margins even in the presence of ongoing deformations. This allows image-based control of tumor resections and can result in the complete removal of lesions with less or no damage to adjacent normal tissues. Intraoperative complications, such as hemorrhage or edema, can be immediately identified and their resolution can be facilitated (Black *et al.*, 1997; Schwartz *et al.*, 1999; Hall *et al.*, 1998, 1999, 2000; Sutherland *et al.*, 1999; Tronnier *et al.*, 1999; Wirtz *et al.*, 1998; Rubino *et al.*, Martin *et al.*, 1999). Craniotomies using MRI guidance are performed routinely, and

lesions treated include intracranial hemorrhages, cysts, as well as malignant and benign brain tumors, cavernous hemangiomas, and arteriovenous malformations (Black *et al.*, 1997; Schwartz *et al.*, 1999). During open surgery the surgeon cannot see beyond the visible surface, so it is very helpful to use intraoperative volumetric imaging to depict the entire operational volume during the intracranial surgery. Today the major benefit of intraoperative imaging is to control tumor resections and to reduce the possibility of residual tumor. More precise definition of target tissue, as well as functional and structural areas to be avoided, with functional MRI and diffusion-weighted images and diffusion tensor representation of the white matter tracts will continue to improve IMRI. Improvements to MR imaging will include faster acquisition times and refinement of imaging sequences for neurosurgical guidance such as continuous imaging (Kacher *et al.*, 2000). Most importantly, in the future, intraoperative guidance may result in major changes in operative approaches and in introducing novel surgical techniques.

In neurosurgery the introduction of real-time image updates has improved localization and targeting and the completeness of tumor resections. The same fundamental approach can also be used in other surgical fields. MRI provides high-sensitivity identification of the margins of breast cancer. The approach that was refined for brain tumor detection and removal can be directly applied for lumpectomy. One of the major challenges in performing lumpectomies is the intraoperative detection of tumor boundaries. Because the breast is less rigid than the brain, the use of preoperative image data is even more restrictive than in the case of brain. In lumpectomy it is highly desirable to use intraoperative image updates to identify margins and to recognize residual tumor.

V. THERMAL ABLATIONS

Originally, image guidance was limited to the deployment of various probes through which the biopsy and/or the treatment was accomplished (*targeting*). Although correct targeting is very important, imaging can continue during the procedure (*monitoring*). The resurgence of local tissue-killing techniques using various forms of energy (chemical, thermal) is justified by increasing capabilities of monitoring and *control* by advanced image-guidance methods (US, CT, MRI).

The unique potential of MRI to detect temperature changes initiated and inspired the evolution of interventional MRI from simple MRI-guided biopsy method to a sophisticated tool to monitor or potentially control thermal ablations (Silverman *et al.*, 1995; Morrison *et al.*, 1998; Matsumoto *et al.*, 1994; Young *et al.*, 1994; Dickinson *et al.*, 1986; Patel *et al.*, 1998; Ishihara *et al.*, 1995; Kuroda *et al.*, 1997; Stollberger *et al.*, 1998; Bertsch *et al.*, 1998). Thermal ablations are effective, minimally invasive methods for tumor treatment if appropriate guidance of the thermal deposition is achieved. For a successful thermal therapy the tumor should be localized and targeted and heated or cooled. The treatment is unsuccessful unless the right temperature range is achieved and maintained for an appropriate time period. In addition, the damage to the adjacent normal tissue must be minimized. These goals cannot be accomplished without image guidance. If the thermal treatment can be combined with temperature monitoring, the full potential of thermal therapy can be demonstrated. The development of temperature-sensitive MRI techniques which improve the efficacy of various thermal treatments can help in the resurgence of thermal ablations. The field is now best characterized as being in an early stage of development and mostly involves the testing of feasibility.

MRI's high sensitivity for localizing tumor margins and the surrounding anatomic structures can be used for targeting. The multiplanar capability helps in trajectory optimization and in correct targeting by various probes. Multiple temperature-sensitive MRI parameters (T1, diffusion, and chemical shift) are relevant for thermal mapping and monitoring. MRI can demonstrate thermally induced changes in diffusion and perfusion and characterize tissue injury.

The biological mechanisms of heat-mediated tissue damage are well known, but the entire spatial extent of the tissue damage can only be demonstrated using volumetric imaging. This requires not only accurate target definition, but also sensitive monitoring. Temperature sensitive MRI monitoring can be used to control the deposition of thermal energy and can detect potential energy spread to the surrounding normal tissue. This way a thermal ablative treatment can be highly effective and safe.

Imaging is an important, but it is not the only factor of image guidance. The integration of imaging and therapy is requisite for the control of interventional manipulations and for optimal energy delivery during thermal ablations. This control can be accomplished only if the characteristics of the imaging systems and the features of the therapy devices are matched and their functional properties are coordinated. Both the temporal and spatial resolution of imaging have to satisfy the requirements for the particular intervention. The time constants of thermal diffusion or the identification of the tip of thermal energy delivery probe (optical fiber, radiofrequency needle, etc.) are equally important to perform a safe and efficient image-guided therapy. These factors all should be seriously considered before an image-guided therapy procedure has been conceived, developed, and implemented.

A. Interstitial Laser Therapy

Interstitial laser therapy (ILT) is a minimally invasive ablative procedure designed for tumor coagulation. ILT uses a laser as a heating source and applies near-infrared emission (such as via a neodymium–yttrium–aluminum garnet, Nd:YAG) laser to deliver energy directly to tissue through optical fibers. The distribution of energy depends on both the optical and thermal properties of the tissue (scattering, absorption, thermal conductivity, and perfusion). Primary optical absorption and subsequent thermal conduction result in irreversible tissue coagulation. This coagulation occurs at and above 60°C.

MRI is well suited for monitoring ILT (Jolesz *et al.*, 1988). The optical fibers generally have small diameters and can pass through thin needles. Therefore, laser treatment is convenient and adaptable for converting a biopsy procedure into a treatment session. The optical fibers and the light itself are fully compatible with MRI, which can provide fast and relatively accurate temperature-sensitive images with appropriate temporal resolution. Soon after the original suggestion of MRI-guided ILT, clinical applications for brain tumor treatment were initiated (Kahn *et al.*, 1998; Bettag *et al.*, 1991; Ascher *et al.*, 1991; Schwabe *et al.*, 1997). The treatments were mostly for malignant gliomas and brain metastases.

Preoperative localization can be complemented by electrophysiological methods and by fast MRI. Monitoring can be accomplished by T1-weighted, or phase-sensitive, MRI. Control by on-line monitoring makes ILT particularly suitable for brain tumors that are located in areas of functional relevance. Soon after the introduction of MRI-guided ILT in neurosurgery, other clinical applications were tested. MRI-guided thermal therapy has become an accepted, minimally invasive treatment option for liver tumors, breast cancer, and head and neck malignancies and more recently for the treatment of uterine fibroids. It is a relatively simple and straightforward method which can be well adapted to the MRI environment (Kettenbach *et al.*, 1998; Vogl *et al.*, 1997; Kahn *et al.*, 1998).

B. Focused Ultrasound Surgery

One of the most attractive approaches for thermal ablation is based on the use of focused ultrasound (FUS). Ultrasound penetrates through soft tissues and can be focused to a few millimeters. The acoustic energy is absorbed and causes temperature elevation with a relatively narrow thermal gradient. It is possible to achieve sharply demarcated target volumes without damaging the adjacent normal tissues. Similar well-controlled focusing of thermal energy cannot be achieved by other heating methods especially without an invasive probe.

The feasibility of MRI-guided FUS to monitor the therapy has been demonstrated (Cline *et al.*, 1994; Chung *et al.*, 1996; Hynynen *et al.*, 1996, 1997, n.d.). MRI represents major advantages over other imaging techniques for targeting, monitoring, and controlling ultrasound exposures. MRI's excellent tissue characterization can be exploited for localization and accurate targeting. Temperature-sensitive MRI sequences can be used to monitor temperature changes and to detect irreversible tissue damage (Young *et al.*, 1994; Dickinson *et al.*, 1986; Patel *et al.*, 1998; Ishihara *et al.*, 1995; Kuroda *et al.*, 1997; Stollberger *et al.*, 1998; Bertsch *et al.*, 1998). The location of the focus can be depicted at low power levels to verify accurate targeting. The tissue changes induced by the sonications can be detected using T1- and T2-weighted MR images. The occlusion of the microvasculature can be detected by the lack of MRI contrast agent uptake. The temperature history of the treated tissue volumes can be used to calculate the biological effect or thermal dose induced by the exposure. In addition, the imaging an be used to monitor normal tissue temperatures for safety.

The feasibility of using a single, focused ultrasound transducer guided by MRI has been demonstrated in clinical treatments of fibroadenomas of the breast (Fig. 8) (Hynynen *et al.*, n.d.). Phased array ultrasound transducer systems can increase the focal volume and reduce the treatment time. The utilization of the phased arrays allow one to make the thermal exposure distribution uniform and use the minimum amount of power to reduce the total treatment time. The results so far indicate that one can aim the ultrasound beam into the tumor accurately through the breast tissue and that the temperature can be elevated enough to coagulate the tumor tissue. These treatments have shown that MRI can detect temperature elevation during the sonication, and thus the basic concept of MRI-monitored ultrasound surgery is valid (Fig. 8).

VI. CRYOABLATION

The MRI signal from frozen water is minimal or absent. This lack of signal has been exploited for control of cryosurgery by monitoring the signal void of the evolving ice ball with standard fast MRI sequences (Silverman *et al.*, n.d.). Using special an MRI-adapted cryosurgery unit, percutaneous treatment of soft tissue tumors (liver, breast, kidney, muscle, and bone) is possible (Fig. 9). The MRI-guided percutaneous cryotherapy approach is feasible because monitoring of the developing iceball in multiple planes is possible. During the freezing process, dynamic MRI demonstrates the slow expansion of the iceball, while thawing images confirm the receding effect. Permanent tissue changes have been clearly identified and can be followed by serial MRI (Silverman *et al.*, n.d.).

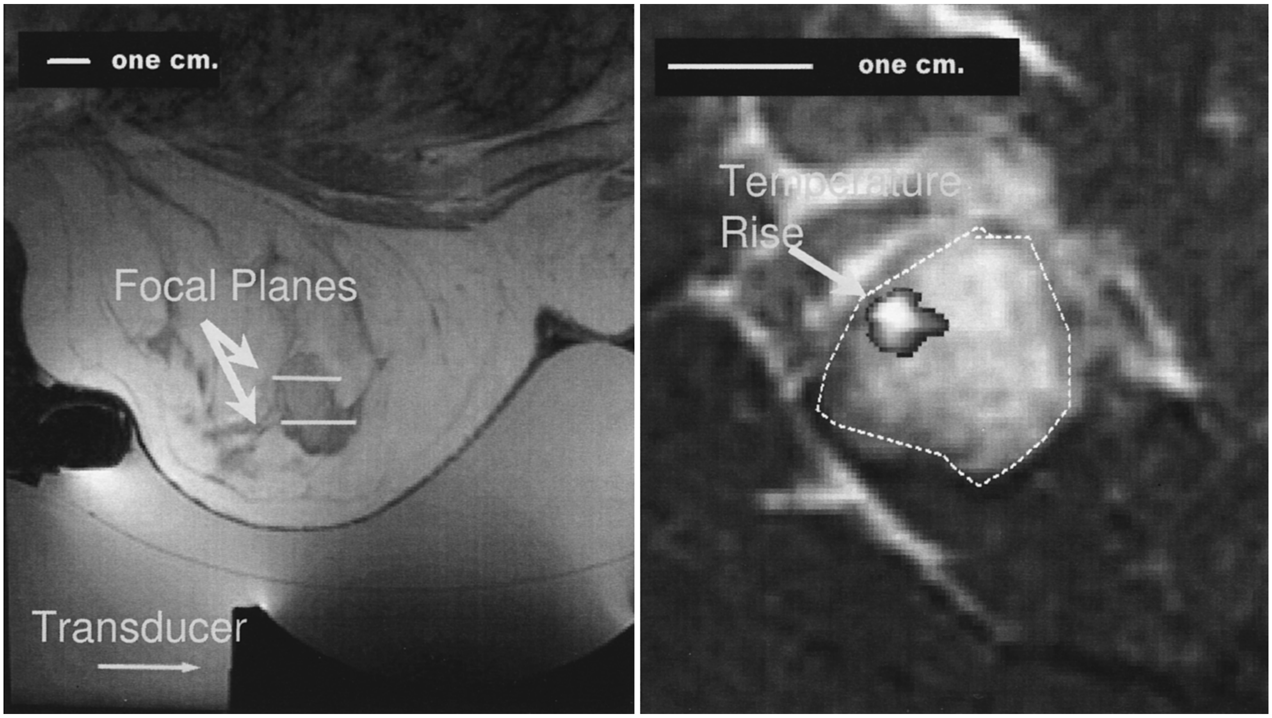


FIGURE 8 Focused ultrasound treatment of breast fibroadenoma. Left: the breast is positioned on a water-filled pillow which provides acoustic coupling between the breast tissue and the transducer. (The transducer is positioned within the MR table.) The focal planes are positioned within the tumor. Right: the outline of the tumor target is defined by the dashed line. The colored area represents temperature rise at the focal point measured by temperature-sensitive MRI sequences.

VII. CONCLUSION

Images contain information which can be used for both diagnosis and therapeutic interventions. These two applications of image-based information, however, cannot be disconnected due to the close interplay between the process of diagnosis and therapy. Nevertheless, there are fundamental differences between the requirements for a diagnostic work-up and an imaging study directed toward a therapeutic procedure. For correct diagnosis specificity has greater significance than sensitivity. For therapy, sensitivity should be a fundamental feature.

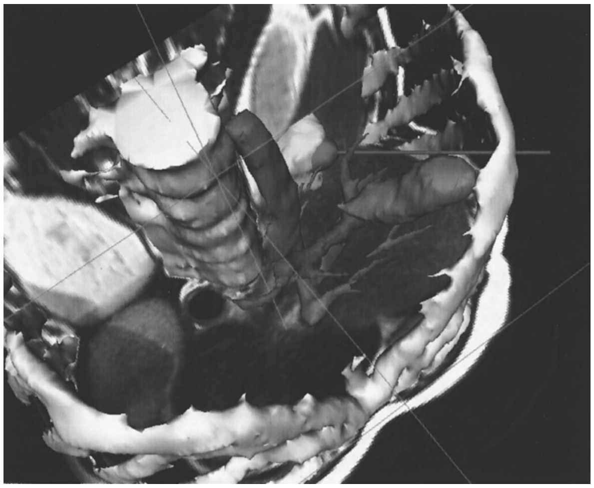


FIGURE 9 Three-dimensional planning of cryoablation of metastatic liver tumor. The yellow represents the tumor, the red is the frozen ice ball at the tip of the needle. The green is the gallbladder and the blue represents portal and hepatic veins.

Images are fundamental in finding optimal access to the target of interventions and in using various targeting methods to define trajectories for instruments. For localization, targeting, and monitoring interventions, all available imaging modalities, but primarily X-ray fluoroscopy, have been exploited. More recently CT, US, and MRI were brought into the operating room environment for intraoperative image guidance. At the same time, with the advance of computerized image processing and visualization tools, image-guidance systems have been introduced for various surgical and radiation oncology applications. These systems use preoperatively acquired images to create anatomic models which provide localization, targeting, and visualization of the three-dimensional anatomy.

Preoperative models, however, should be modified as the procedure progresses and the anatomy changes. The only adequate solution to detect physiologic motion, displacements, or deformations is intraoperative or intraprocedural imaging. Monitoring of dynamic changes not restricted to follow motions but a variety of other functional or physical parameters may be altered or modified during interventional or surgical procedures. Although the primary goal of monitoring is to follow and update anatomic changes in position, other types of dynamic information (flow, perfusion, cortical function, etc.) can also be extremely useful.

MRI's high sensitivity and superb tissue characterization explain its success in intraoperative monitoring for brain tumor surgery or for breast cancer excision. During these open surgeries the tissue planes significantly shift and deform, therefore the original template of the tumor obtained by preoperative imaging or by images acquired at the beginning of the surgery become useless as the procedure progresses. Real-time monitoring or frequent volumetric update is necessary to redefine the location of the tumor, its surrounding, the advance of surgery, and the presence or absence of any local tissue response or threatening complications. New intraoperative monitoring and interactive imaging methods are intended to convert interventional radiology and minimally invasive surgical techniques, including thermal ablations and brachytherapy, to genuine image-guided therapy where the role of image-based information is crucial and strategically important during every phase of the entire procedure.

ACKNOWLEDGMENTS

The illustrations were provided by Ron Kikinis, M.D., Stuart G. Silverman, M.D., Peter McL. Black, M.D., Ph.D., Arya Nabavi, M.D., Clare M. Tempany, M.D., Anthony V. D'Amico, M.D., Ph.D., Torsten Butz, Kullervo Hynynen, Ph.D., Robert A. Cormack, Ph.D., Andreas G. Schreyer, M.D., Nobi Hata, Ph.D., Cynthia Wible, Ph.D., and Fatma Ozlen, M.D., from the Image Guided Therapy Program, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts.

SEE ALSO THE FOLLOWING ARTICLES

Image Processing • Imaging Optics • Lasers • Magnetic Resonance in Medicine • Optical Fiber Techniques for Medical Applications • Phamaceuticals, Controlled Release of • X-Ray Analysis

BIBLIOGRAPHY

Ascher, P. W., Justich, E., and Schrottner, O. (1991). "Interstitial thermotherapy of central brain tumors with the Nd:YAG laser under real-time monitoring by MRI," *J. Clin. Laser Med. Surg.* **9,** 79–83.

Barnett, G. H., Kormos, D. W., Steiner, C. P., and Weisenberger, J. (1993). "Use of a frameless, armless stereotactic wand for brain tumor localization with two-dimensional and three-dimensional neuroimaging," *Neurosurgery* **33,** 674–678.

Bertsch, F., Mattner, J., Stehling, M. K., Muller-Lisse, U., Peller, M., Loeffler, R., Weber, J., Messmer, K., Wilmanns, W., Issels, R., and Reiser, M. (1998). "Non-invasive temperature mapping using MRI: Comparison of two methods based on chemical shift and T1-relaxation," *Magn. Reson. Imaging* **16,** 393–404.

Bettag, M., Ulrich, F., Schober, R., Furst, G., Langen, K. J., Sabel, M., and Kiwit, J. C. (1991). "Stereotactic laser therapy in cerebral gliomas," *Acta. Neurochir. Suppl.* **52,** 81–83.

Black, P. McL., Moriarty, T., Alexander, E., Stieg, P., Woodard, E. J., Gleason, P. L., Martin, C. H., Kikinis, R., Schwartz, R. B., and Jolesz, F. A. (1997). "Development and implementation of intraoperative magnetic resonance imaging and its neurosurgical applications," *Neurosurgery* **41,** 831–843.

Chung, A., Hynynen, K., Cline, H. E., Colucci, V., Oshio, K., and Jolesz, F. A. (1996). "Optimization of spoiled gradient-echo phase imaging for *in vivo* localization of focused ultrasound beam," *Magn. Reson. Med.* **36,** 745–752.

Cline, H. E., Lorensen, W. E., Kikinis, R., and Jolesz, F. A. (1990). "Three-dimensional segmentation of MR images of the head using probability and connectivity," *J. Comput. Assist. Tomogr.* **14,** 1037–1045.

Cline, H. E., Hynynen, K., Hardy, C. J., Watkins, R. D., Schenk, J. F., and Jolesz, F. A. (1994). "MR temperature mapping of focused ultrasound surgery," *Magn. Reson. Med.* **31,** 628–636.

Cotin, S., Delingette, H., and Ayache, N. (1999). "Real-time elastic deformations of soft tissues for surgery simulation," *IEEE Trans. Visualization Computer Graphics* **5,** 62–73.

Cover, S., Ezquerra, N., O'Brian, J., Rowe, R., Gadacz, T., and Palm, E. (1993). "Interactively deformable models for surgery simulation," *IEEE Computer Graphics Appl.* **13,** 68–75.

D'Amico, A. V., Cormack, R., Tempany, C. M., Kumar, S., Topulos, G., Kooy, H. M., and Coleman, C. N. (1998). "Real-time magnetic resonance image-guided interstitial brachytherapy in the treatment of select patients with clinically localized prostate cancer," *Int. J. Radial. Oncol. Biol. Phys.* **42,** 507–515.

Dickinson, R. J., Hall, A. S., Hind, A. J., and Young, I. R. (1986). "Measurement of changes in tissue temperature using MR imaging," *J. Comput. Assist. Tomogr.* **10,** 468–472.

Fried, M. P., Hsu, L., and Jolesz, F. A. (1998). "Interactive magnetic resonance imaging-guided biopsy in the head and neck: Initial patient experience," *Laryngoscope* **1998,** 108–148.

Galloway, R. L., Maciunas, R. J., and Edwards, C. A. (1992). "Interactive image-guided neurosurgery," *IEEE Trans. Biomed. Eng.* **39,** 1226–1231.

Gibson, S., Fyock, C., Grimson, E., Kanade, T., Kikinis, R., Lauer, H., McKenzie, N., Mor, A., Nakajima, S., Ohkami, H., Osborne, R., Samosky, J., and Sawada, A. (1998). "Volumetric object modeling for surgical simulation," *Med. Image Anal.* **2,** 121–132.

Grimson, W. E. L., Ettinger, G. J., White, S. J., Lozano-Perez, T., Wells, W. M., and Kikinis, R. (1996). "An automatic registration method for frameless stereotaxy, image guided surgery, and enhanced reality visualization," *IEEE Trans. Med. Imaging* **15,** 129–140.

Hall, W. A., Martin, A. J., Liu, H., Pozza, C. H., Casey, S. O., Michel, E., Nussbaum, E. S., Maxwell, R. E., and Truwit, C. (1998). "High-field strength interventional magnetic resonance imaging for pediatric neurosurgery," *Pediatr. Neurosurg.* **29,** 253–259.

Hall, W. A., Martin, A. J., Liu, H., Nussbaum, E. S., Maxwell, R. E., and

Truwit, C. L. (1999). "Brain biopsy using high-field strength interventional magnetic resonance imaging," *Neurosurgery* **44,** 807–814.

Hall, W. A., Liu, H., Martin, A. J., Pozza, C. H., Maxwell, R. E., and Truwit, C. L. (2000). "Safety, efficacy, and functionality of high-field strength interventional magnetic resonance imaging for neurosurgery," *Neurosurgery* **46,** 632–642.

Hata, N., Dohi, T., Warfield, S., Wells, W. M., Kikinis, R., and Jolesz, F. A. (1998). "Multimodality deformable registration of pre- and intraoperative images for MRI-guided brain surgery," *In* "First International Conference on Medical Image, Computing and Computer-assisted Interventions," pp. 1067–1074, Springer-Verlag, Berlin.

Heilbrun, M. P., McDonald, P., Wiker, C., Koehler, S., and Peters, W. (1992). "Stereotactic localization and guidance using a machine vision technique," *Stereotact. Funct. Neurosurg.* **58,** 94–98.

Held, K., Kops, E. R., Krause, B., and Wells, W. (1996). "Markov random field segmentation of brain MR images," *IEEE Trans. Med. Imaging* **16,** 878–887.

Hsu, L., Fried, M. P., and Jolesz, F. A. (1998). "MR-guided endoscopic sinus surgery," *AJNR Am. J. Neuroradiol.* **19,** 1235–1240.

Hynynen, K., Freund, W. R., Cline, H. E., Chung, A. H., Watkins, R. D., Vetro, J. P., and Jolesz, F. A. (1996). "Non-invasive MR imaging-monitored ultrasound surgery method," *RadioGraphics* **16,** 185–195.

Hynynen, K., Vykhodtseva, N. I., Chung, A. H., Sorrentino, V., Colucci, V., and Jolesz, F. A. (1997). "Thermal effects of focused ultrasound on the brain: Determination with MR imaging," *Radiology* **204,** 247–253.

Hynynen, K., Pomeroy, O., Smith, D., Huber, P., McDannold, N. J., Kettenbach, J., Baum, J., Singer, S., and Jolesz, F. A. (2001). "MRI guided focused ultrasound surgery (FUS) of fibroadenomas in the breast," *Radiology* **219**(1), 176–185.

Ishihara, Y., Calderon, A., Watanabe, H., Okamoto, K., Suzuki, Y., and Kuroda, K. (1995). "A precise and fast temperature mapping using water proton chemical shift," *Magn. Reson. Med.* **34,** 814–823.

Jolesz, F. A. (1997). "Image-guided procedures and the operating room of the future," *Radiology* **204,** 601–612.

Jolesz, F. A. (1998). "Interventional and intraoperative MRI: A general overview of the field," *J. Magn. Reson. Imaging* **8,** 3–7.

Jolesz, F. A., and Blumenfeld, M. (1994). "Interventional use of magnetic resonance imaging," *Magn. Reson.* **10,** 85–96.

Jolesz, F. A., and Shtern, F. (1992). "The operating room of the future. Report of the National Cancer Institute Workshop, Imaging Guided Stereotactic Tumor Diagnosis and Treatment," *Invest. Radiol.* **27,** 326–328.

Jolesz, F. A., Bleier, A. R., Jakab, P., Ruenzel, P. W., Huttl, K., and Jako, G. P. (1988). "MR imaging of lasler-tissue interactions," *Radiology* **168,** 249–253.

Kacher, D. F., Maier, S. E., Mamata, H., Mamata, Y., Nabavi, A., and Jolesz, F. A. (2001). "Motion robust imaging for continuous intraoperative MRI," *J. Magn. Reson. Imaging* **13**(1), 158–161.

Kahn, T., Harth, T., Kiwit, J. C., Schwarzmaier, H. J., Wald, C., and Modder, U. (1998). "*In vivo* MRI thermometry using a phase-sensitive sequence: Preliminary experience during MRI-guided laser-induced interstitial thermotherapy of brain tumors," *J. Magn. Reson. Imaging* **8,** 160–164.

Kettenbach, J., Silverman, S. G., Hata, N., Kuroda, K., Saiviroonporn, P., Zientara, G. P., Morrison, P. R., Hushek, S. G., McL. Black, P., Kikinis, R., and Jolesz, F. A. (1998). "Monitoring and visualization techniques for MR-guided laser ablations in open MR-system," *J. Magn. Reson. Imaging* **8,** 933–943.

Kikinis, R., Gleason, P. L., Moriarty, T. M., Moore, M. R., Alexander, E., Stieg, P. E., Matsumae, M., Lorensen, W. E., Cline, H. E., Black, P. M., and Jolesz, F. A. (1996). "Computer assisted interactive three-dimensional planning for neurosurgical procedures," *Neurosurgery* **38,** 640–651.

Koivukangas, Y., Louhisalmi, J., Alakuijala, H., and Oikarinen, J. (1993). "Ultrasound-controlled neuronavigator-guided brain surgery," *J. Neurosurg.* **79,** 36–42.

Kuroda, K., Oshio, K., Chung, A. H., Hynynen, K., and Jolesz, F. A. (1997). "Temperature mapping using the water proton chemical shift: A chemical shift selective phase mapping method," *Magn. Reson. Med.* **38,** 845–851.

Laborde, G., Gilsbach, J., Harders, A., Klimek, L., Moesges, R., and Krybus, W. (1992). "Computer assisted localizer for planning of surgery and intra-operative orientation," *Acta. Neurochir.* **119,** 166–170.

Maciunas, R. J., Galloway, R. L., Fitzpatrick, J. M., Mandava, V. R., Edwards, C. A., and Allen, G. S. (1992). "A universal system for interactive image-directed neurosurgery," *Stereotact. Funct. Neurosurg.* **58,** 108–113.

Martin, A. J., and van Vaals, J. J. (1999). "High-field interventional MR system for neurologic applications," *In* "Interventional MRI" (R. B. Lufkin, ed.), pp. 35–49, Mosby, St. Louis, MO.

Matsumoto, R., Mulkern, R. V., Hushek, S. G., and Jolesz, F. A. (1994). "Tissue temperature monitoring for thermal interventional therapy: Comparison of T1-weighted MR sequences," *J. Magn. Reson. Imaging* **4,** 65–70.

Morrison, P. R., Jolesz, F. A., Charous, D., Mulkern, R. V., Hushek, S. G., Margolis, R., and Fried, M. P. (1998). "MRI of laser-induced interstitial thermal injury in an *in vivo* animal liver model with histologic correlation," *J. Magn. Reson. Imaging* **8,** 57–63.

Nakajima, S., Atsumi, H., Bhalerao, A. H., Jolesz, F. A., Kikinis, R., Yoshimine, T., Moriarty, T. M., and Stieg, P. E., (1997). "Computer-assisted surgical planning for cerebrovascular neurosurgery," *Neurosurgery* **41,** 403–410.

Patel, K. C., Duerk, J. L., Zhang, Q., Chung, Y. C., Williams, M., Kaczynski, K., Wendt, and M., and Lewin, J. S. (1998). "Methods for providing probe position and temperature information of MR images during interventional procedures," *IEEE Trans. Med. Imaging* **17,** 794–802.

Pelizzari, C. A., Chen, G. T., Spelbring, D. R., Weichselbaum, R. R., and Chen, C. T. (1989). "Accurate three-dimensional registration of CT, PET, and/or MR images of the brain," *J. Comput. Assist. Tomogr.* **23,** 20–26.

Rubino, G. R., Farahani, K., McGill, D., Van de Wiele, B., Villablanca, J. P., and Wang-Mathieson, A. (2000). "Magnetic resonance imaging-guided neurosurgery in the magnetic fringe fields: The next step in neuronavigation," *Neurosurgery* **46,** 643–654.

Schenk, J. F., Jolesz, F. A., Roemer, P. M., Cline, H. E., Lorensen, W. E., Vosburgh, K. G., and Kikinis, R. (1995). "Superconducting open-configuration MR imaging system for image guided therapy," *Radiology* **1995,** 805–814.

Schwabe, B., Kahn, T., Harth, T., Ulrich, F., and Schwarzmaier, H. J. (1997). Laser-induced thermal lesions in the human brain: Short- and long-term appearance on MRI," *J. Comput. Assist. Tomogr.* **21,** 818–825.

Schwartz, R. B., Hsu, L., Wong, T. Z., Kacher, D. F., Zamani, A. A., Black, P. M., Alexander, E., 3rd, Stieg, P. E., Moriarty, T. M., Martin, C. A., Kikinis, R., and Jolesz, F. A. (1999). "Intraoperative MR imaging guidance for intracranial neurosurgery: Experience with the first 200 cases," *Radiology* **211,** 477–488.

Silverman, S. G., Collick, B. D., Figueira, M. R., Khorasani, R., Adams, D. F., Newmans, R. W., Topulos, G. P., and Jolesz, F. A. (1995). "Interactive MR guided biopsy in an open configuration MR imaging system," *Radiology* **197,** 175–181.

Silverman, S. G., Jolesz, F. A., Newman, R. W., et al. (1998). "Design

and implementation of an interventional MR imaging suite," *Am. J. Roentgenol.* **168,** 1465–1471.

Silverman, S. G., Tuncali, K., Adams, D. S., Kahcer, D., Morrison, P. R., Granter, S. R., Osteen, R. T., and Jolesz, F. A. (2000). "Percutaneous MRI-guided cryotherapy of livertumores: A preliminary report of feasibility and safety," *Radiology*, **217**(3), 657–664.

Smith, K. R., Frank, K. J., and Bucholz, R. D. (1994). "The NeuroStation—a highly accurate, minimally invasive solution to frameless stereotactic neurosurgery," *Comput. Med. Imaging Graph* **18**(4), 247–256.

Stollberger, R., Ascher, P. W., Huber, D., Renhart, W., Radner, H., and Ebner, F. (1998). "Temperature monitoring of interstitial thermal tissue coagulation using MR phase images," *J. Magn. Reson. Imaging* **8,** 188–196.

Sutherland, G. R., Kaibara, T., Louw, D., Hoult, D. I., Tomanek, B., and Saunders, J. (1999). "A mobile high-field magnetic resonance system for neurosurgery," *J. Neurosurg.* **91,** 804–813.

Tronnier, V., Staubert, A., Wirtz, R., Knauth, M., Bonsanto, M., and Kunze, S. (1999). "MRI-guided brain biopsies using a 0.2 Tesla open magnet," *Minim. Invasive Neurosurg.* **42,** 118–122.

Vogl, T. J., Mack, M. G., Straub, R., Roggan, A., and Felix, R. (1997). "Magnetic resonance imaging—Guided abdominal interventional radiology: Laser-induced thermotherapy of liver metastases," *Endoscopy* **29,** 577–583.

Warfield, S. K., Jolesz, F., and Kikinis, R. (1998). "A high performance computing approach to the registration of medical imaging data," *Parallel Computing* **24,** 1345–1368.

Wells, W. M., Kikinis, R., Grimson, W. E. L., and Jolesz, F. A. (1996a). "Adaptive segmentation of MRI data," *IEEE Trans. Med. Imaging* **15,** 429–442.

Wells, W. M., Viola, P., Atsumi, H., Nakajima, S., and Kikinis, R. (1996b). "Multi-modal volume registration by maximization of mutual information," *Med. Image Anal.* **1,** 35–51.

Wirtz, C. R., Tronnier, V. M., Albert, F. K., Knauth, M., Bonsanto, M. M., Staubert, A., Pastyr, O., and Kunze, S. (1998). "Modified headholder and operating table for intra-operative MRI in neurosurgery," *Neurol. Res.* **20,** 658–661.

Young, I. R., Hand, J. W., Oatridge, A., Prior, M. W., and Forse, G. R. (1994). "Further observations on the measurement of tissue T1 to monitor temperature *in vivo* by MRI," *Magan. Reson. Med.* **31,** 342–345.

Zamorano, L. J., Nolte, L., Kadi, A. M., and Jiang, Z. (1993). "Interactive intraoperative localization using an infrared-based system," *Neurol. Res.* **15,** 290–298.

Zinreich, S. J., Tebo, S. A., Long, D. M., Brem, H., Mattox, D. E., Loury, M. E., vander Kolk, C. A., Koch, W. M., Kennedy, D. W., and Bryan, R. N. (1993). "Frameless stereotaxic integration of CT imaging data: Accuracy and initial applications," *Radiology* **188,** 735–742.

Mammalian Cell Culture

Bryan Griffiths
SC&P

Florian Wurm
Swiss Federal Institute of Technology Lausanne

I. Introduction
II. The Cell
III. The Culture System
IV. The Culture Production Process
V. Cell Products
VI. Applications of Cell Culture
VII. Conclusion

GLOSSARY

Amplification The generation of multiple copies of one or more genes in a cell or nucleus.

Anchorage-dependent cell (ADC) Cells that will grow, or survive, only when attached to a surface (e.g., glass or plastic).

Calcium-phosphate/DNA co-precipitation The most popular method for the preparation of DNA that is to be transfected into mammalian cells.

Cell hybridization The fusion of two or more dissimilar cells leading to the formation of a synkaryon (hybrid with fused nuclei from original cells).

Cell line A cell line arises from a primary culture at the time of the first successful subculture and implies that cultures consist of cell lineages originally present in the primary culture.

Clone A population of cells derived from a single cell by mitosis.

Continuous cell culture A culture apparently capable of an unlimited number of population doublings (often referred to as *immortal*).

Dihydrofolate reductase (DHFR) An enzyme that reduces dihydrofolate to tetrahydrofolate.

Immortalization The attainment by a finite cell culture of the attributes of a continuous cell line.

Methotrexate A folic acid antagonist that binds to, and inhibits, DHFR (see above).

Recombinant protein A protein produced in a cell that has been genetically engineered (e.g., by transfection).

Subculture (=passage) The transfer of cells, with or without dilution, from one culture vessel to another.

Suspension cell Cells, that either singly or in aggregates multiply while suspended in liquid medium.

Transfection The transfer of naked DNA into cells in culture.

Transformation An inheritable change to cells in culture, either intrinsically or by treatment with carcinogens, oncogenic viruses, irradiation, or transfection with oncogenes, leading to the acquisition of altered properties (e.g., neoplatic, proliferative, antigenic).

Tissue plasminogen activator (tPA) Plasminogen activators catalyze the conversion of plasminogen to the

active fibrinolytic enzyme plasmin and are used for dissolving blood clots.

Transgenic animals Animals that have incorporated foreign DNA heritably.

CELL CULTURE refers to the ability to grow cells derived from the whole organism as either discrete cells or as small fragments of tissue *in vitro*. Cell culture has become a very important technology in a wide range of life science applications. It allows the study of cell growth and control, differentiation, genetics, and many diseases including cancer. In addition, it is extensively used for the manufacture of a large number of biological products including vaccines, hormones, immunologicals, and blood factors and for tissue engineering and gene therapy.

I. INTRODUCTION

The initial aims of cell and tissue culture were to study specialized cell behavior and function *in vitro*. However, these aims could not be realized because only the ubiquitous dedifferentiated cell, or cells transformed by carcinogens, survived in culture. Although mammalian cells have been grown *in vitro* since before 1907, the factor that gave impetus to their current widespread laboratory and industrial use was the discovery by Enders in 1949 that human pathogenic viruses could be grown in cell cultures. Prior to this, viruses could only be grown in living tissue, thus vaccine production used living organisms such as the embryonic chicken. The use of cultured cells to grow viruses opened up the possibility of a less expensive, easier, biologically safer, more controllable (and reproducible), and larger scale method for vaccine manufacture. Following this demonstration by Enders, it took only 5 years before the first cell-based vaccine was licensed for clinical use (the Salk polio vaccine in primary monkey kidney cells in 1954). This opened up 20 years of continuous development of human and veterinary viral vaccines and created the need for industrial-scale cell-culture processes. The most effective large-scale process developed during this period was for foot and mouth disease virus (FMDV) based on suspension culture of BHK cells. Developments in processes for human vaccines were less dramatic due to the need to use biologically safe cell lines. This meant the use of human diploid cell lines, such as WI-38 and MRC-5, which unfortunately, due to their normality (i.e., they behave as cells *in vivo* without tumorgenic transformation), only grow attached to a substrate (anchorage dependent) and only reach low cell densities. The wide range of animal cell reactors available is partly due to the dual development of systems for anchorage-dependent cells (ADC) and free-suspension cells. ADC were grown in small culture vessels and a production batch constituted hundreds, even thousands, of replicate cultures (i.e., a multiple batch process). The need for a unit batch process (one large culture vessel), such as the fermenters used for suspension cells, saw the development of a wide range of novel culture reactors, but the significant breakthrough came with the microcarrier system. This procedure developed in 1967 by van Wezel allowed cells to grow attached to small (200-micron) spheres which were stirred in a large tank fermenter analogous to suspension cells. This was the first successful large-scale unit process for ADC.

Vaccines were the dominant product until the 1970s, but changes in regulatory and licensing procedures then allowed cells from sources other than normal tissues to be used for human medicinal products. This came about during the development of a production process for human interferon proteins using a cancer cell line, Namalva, by the company Wellcome. Their pioneering work established the safety criteria, and thus acceptance, for using non-normal (heteroploid, transformed, or tumor-derived) cell lines and the feasibility of scaling-up an industrial cell culture process to 8000 L. Cell culture then entered a new, or modern, phase where a wide variety of cell products (Table I) is produced from a range of cell types (see Section II). Two of these can be highlighted as being significant milestones. First, the production of monoclonal antibodies from hybridoma cells (the fusion of a normal antibody-producing cell and a hemopoitic cancer cell) which has given rise to hundreds of new products. Second, the development of recombinant tPA by Genentech which gave rise to the first genetically engineered clinical product from cell cultures. Currently, applications are widening to include the cell itself as a product in tissue engineering and organ replacement and for gene therapy.

Technological advances have obviously driven the development of animal cell biotechnology from 1954 to the present day, but the main influencing factor has been the safety of the end product. Regulatory bodies such as the World Health Organization (WHO), the U.S. Food and Drug Administration (FDA), and others have set down at all stages of the process acceptable standards for cell products, and these have had to safeguard against both known and perceived hazards such as transforming viruses, disease agents, carcinogenic and immunologically damaging molecules, and, more recently, prions.

In this chapter, emphasis is placed on scale-up because of the relatively low biomass productivity of natural products from animal cells compared to bacteria and the need to introduce more efficient and economical industrial processes to meet the production requirements of recombinant proteins. However, of equal importance

TABLE I Mammalian Cell Products

Native products	Product name (Year of license)
Human vaccines	Polio (1954), measles (1963), rabies (1964), mumps (1969)
Veterinary vaccines	FMDV, rabies, Marek's, pseudorabies, BVD, Louping ill, bluetongue, avian influenza, canine distemper
Interferon	
Recombinant products	
Monoclonal antibodies	OKT3/Orthoclone (1987), Centoxin (1990), Reopro (1994), Myoscint (1989), Oncoscint (1990)
tPA	Activase/Actilyse (1987)
EPO	Epogen/Procrit/Eprex (1989), Epogin/Recormon (1990)
hGH	Saizen (1989)
HBsAg	GenHevac B Pasteur (1989), HBGamma (1990)
Interferon	Roferon(1991)
G-CSF	Granocyte (1991), Neupogen (1991)
Blood factor VIII	Recombinate (1992), Kogenate (1993)
Dnase I	Pulmozyme (1993)
Glucocerebrosiduse	Cerezyme (1994)
FSH	Gonal-F (1995)
rDNA products in development/clinical trial	
HIV vaccines (gp120, gp160, CD4)	
Herpes simplex vaccines (gB,gD)	
Chimeric Mabs (her2, CD4, TNFα, CD20, Cd18, TAC, leukointegrin, CF54, RSV)	
In vitro diagnostic Mabs (over 200)	
Others (TSH, TNF, M-CSF, IL-6, IL-1)	
Tissue engineering and replacement (e.g., skin, artificial liver, kidneys)	

has been the development of more efficient culture media (particularly the identification of specific growth factors that have allowed the introduction of serum-free and low-protein media and thus the growth of specialized rather than undifferentiated cells), followed by cell fusion techniques (e.g., hybridomas) and recombinant DNA technology to allow product expression from fast-growing undifferentiated cells and to enhance productivity.

II. THE CELL

A. Cell Types

A cell culture is usually initiated by the explant technique (allowing cells to migrate out of a tissue fragment to form a culture of individual cells) or by mechanically and enzymically breaking down an organ/tissue into single cells which are then plated out as a primary culture. A complex medium of amino acids, vitamins, salts, glucose, and fetal calf serum (or a range of growth factors normally present in serum), buffered at pH 7.0 to 7.4, and incubated at 37°C is necessary to isolate and cultivate cells.

Cells can be grown as:

1. *Organ culture*: Short-term culture of functional tissue (e.g., tissue slices).
2. *Primary cells*: Short-term culture of single cells isolated directly from tissues, usually by an enzymic treatment, and allowed to grow and divide until they are ready for subculturing into daughter cultures as a cell line. However, for many applications, cells are only used in the primary culture as they still retain some *in vitro* specialized characteristics (e.g., chick embryo fibroblasts and monkey kidney cells) for vaccine manufacture.
3. *Finite cell lines*: Cells derived from normal tissue via a primary culture which can replicate and undergo limited subculture until they become senescent (e.g., WI-38, MRC-5) or immortalized (by chemical carcinogens, transforming viruses, hybridization, or genetic engineering).
4. *Continuous cell lines*: Cells that have an indefinite subculture potential derived from tumour tissue (e.g., HeLa) or that have undergone immortalization *in vitro* (e.g., L929 cells).

Finite cell lines have been extensively used in vaccine production but they are limited in cell type (usually

TABLE II Examples of Commonly Used Mammalian Cell Lines

Cell	Source	Morphology	Application
Finite cell lines			
MRC-5	Human lung	Fibroblastic	Virus studies, vaccine production, and aging studies
WI-38			
Continuous cell lines			
HeLa	Human cervical carcinoma	Epithelial	Virology
L929	Mouse connective tissue	Fibroblastic	General studies
Vero	Monkey kidney	Fibroblastic	Virology and vaccines
BHK21	Syrian hamster	Fibroblastic	FMDV vaccine
CHO, CHO dhfr–	Chinese hamster ovary	Epithelial	Genetics, recombinant proteins
MDCK	Dog kidney	Epithelial	Virology
3T3	Mouse embryo	Fibroblastic	IGF-1 production
COS	African green monkey kidney	Fibroblastic	SV40 viruses
NS0, Sp2/0	Mouse myeloma	Spherical	Fusion partners to form hybridomas for monoclonal antibodies
J558L	Mouse BALB/C myeloma	Spherical	IgA secretion
Hybridoma	Hybrid of myeloma and plasma cell partner	Spherical	Mab production
Namalwa	Human tumor	Spherical	Interferon production
C6	Rat brain glial tumor	Epithelial	Neurotoxicity studies
GH3	Rat pituitary tumor	Epithelial	Hormone studies
293	Adenovirus transformed HEK	Epithelial	Gene therapy
X CRE/Y, CRIP	Mouse NIH3/Moloney leukemia		
		Spherical	Retrovirus packaging

fibroblasts) and lifespan (typically 50 to 60 population doublings for MRC-5 cells). Continuous cell lines are the most widely used cell materials but the cell is usually dedifferentiated.

A cell line is perpetuated by the subculture technique. For anchorage-dependent cells attached to a substrate, this entails detachment from the substrate by an enzyme (usually 0.25% trypsin) followed by dilution in fresh medium in a 1:2 to 1:20 ratio, depending upon cell type. Cells growing in free suspension are concentrated (by gentle centrifugation) and redistributed into fresh medium.

B. Cell Lines

A very wide selection of cell lines is available from culture collections (e.g., ATCC, ECACC, Riken) and examples are given in Table II of the more widely used ones. For more information, see culture collection catalogs.

C. Hybrid Cell Lines

1. Hybridomas

Hybridoma cell lines are biologically unique, since they have their origin in an experimental setup that sets out to unify two cellular characteristics into one. The first one is the feature of the continual production of a specific antibody (monoclonal) by an individual B-cell derived from an immunized animal (mouse, rat, or human). The second is the feature of unlimited life span, a characteristic of tumor cells. Hybridomas are created by the fusion, through a number of techniques, of a tumor cell line, usually of mouse origin with B-lymphocytes derived from an immunized mouse. The original experiment by Koehler and Milstein in 1975 demonstrated that differentiated mouse B-cells producing a specific type of antibody constitutively could be coerced to grow *in vitro* and scaled-up when fused with cells from an immortal tumor cell line (myeloma). Originally, the fusion partner for mouse B-lymphocytes was a mouse myeloma cell line, P3X63AG8, but today more popular as a fusion partner is the cell line SP2/0, which is a recloned product of a fusion experiment between P3X63AG8 and mouse spleen cell. The fusion of cells of the two populations can be induced by the treatment with a high concentration of polyethylene glycol (PEG) solution (in medium). PEG causes aggregation of membrane-bound proteins, resulting in the outer cell membranes being reduced in proteins—membrane-bound proteins interfere with the fusion of lipid bilayers. Two cell membranes that come into close contact with each other will fuse into one, eventually creating a single cell containing the two individual genomes.

Hybrid cell lines carry, due to their particular type of creation, the chromosome sets of the two partner cells that were at the origin of the fusion. One of the two partners is a cell of immortal character (usually a tumor-derived cell line), and these cells are generally known to exhibit a high degree of cytogenetic instability, a feature that is transferred into the fusion product. A disadvantage of hybridoma cell lines, in comparison with, for example, transfected Chinese hamster ovary (CHO) cell lines expressing an antibody, is an inherent (productivity) instability. This instability is due to the decline and eventual overall loss of cells in the cell population expressing the desired antibody and an overgrowth of nonproducing clones in the cell population. In order to compensate for this disadvantage, single-cell cloning and expansion from "young" hybridomas are frequently necessary while limiting the subcultivation of cell populations to short time frames. To obtain a reasonable number of cell lines expressing the desired antibody, several cloning steps with candidate fusion product cells have to be executed.

Monoclonal antibodies derived from murine hybridomas have been used in a variety of clinical applications, but some of the early promise could not be realized due to human anti-mouse immune reactions induced in treated patients. The concept of fusion of two biological systems for the synthesis of a specific monoclonal antibody has also been used for the creation of murine–human heterohybridomas and for human–human hybridomas. In all these cases, because of the use of cells derived from blood sources, a very important characteristic has been "inherited" that significantly facilitates the scale-up of the cell substrate to very large cell number—the growth of single cells in free suspension.

2. Heterokayons

Somatic cell fusion using Sendai virus or polyethylene glycol, and a selective medium to prevent the growth of the parent (non-hybrid) cells, was a valuable tool for studying somatic cell genetics in the 1960s and 1970s. The development of hybridoma technology used these techniques, but currently recombinant genetic techniques have largely replaced the classical fusion techniques. However, cell fusion is still used to study fundamental genetics and physiology, as well as the more publicized monoclonal antibody technology.

D. Recombinant Cell Lines

1. Nonviral DNA Transfer Vehicles for Mammalian Cells

The transfer of genes of interest into mammalian cells was greatly facilitated by the development of methods that complex DNA molecules into a compacted structure, usually together with agents that compensate its negative charge. The most popular of the complexing agents is calcium phosphate, which can be precipitated in the presence of DNA and generates microscopic particles (up to 3 μm in size, but preferably smaller than 0.3 μm at the time of interaction with the cells). Small calcium phosphate DNA co-precipitates (smaller in diameter than 0.3 μm) can carry as much as 40% of their entire mass as DNA and individual 0.5-μm particles can carry as many as 4500 copies of a standard-sized plasmid DNA. More recently, other lipid- or polymer-based DNA delivery vehicles have been developed, most of them provided now in the form of commercial kits. As with calcium phosphate DNA co-precipitates, the majority of these polymer- or lipid-based vehicles deliver DNA into endosome compartments of the cells. It is assumed that the transfer across the cellular membrane occurs by endocytosis. The different steps of transport and processing of DNA to finally achieve expression from the DNA sequences integrated into the genome (see Fig. 1) are poorly understood. They are a subject of intense research due to increased interest in gene therapy, in which genes need to be delivered to the nucleus of many individual cells of the patient.

2. Transfer of Genes of Interest into Mammalian Cells

The generation of recombinant cell lines (i.e., cell populations that have incorporated into one or more of their chromosomes segments of DNA provided experimentally through viral or nonviral vectors) has been motivated by the desire to understand phenomena of gene expression, gene control, and gene regulation. During the early 1980s, mammalian cells and, in particular, CHO cells with a double mutation/deletion of the dihydrofolate reductase (DHFR) gene locus became a rather convenient substrate for the introduction of genes. The availability of these mutant CHO cells allowed identification, in selective media lacking certain precursors for nucleotide synthesis, of clones of cells that express the DHFR gene. The DHFR gene can be provided by transfection with a plasmid vector into which another gene of interest has been cloned or to which a second plasmid containing that gene of interest was added for the preparation of the transfection cocktail (Fig. 2). In both cases, emerging cells after selection contain the DHFR gene and the gene of interest in their chromosomes, usually genetically closely linked to each other. The purpose of most of these experiments has been and still is to obtain cell lines that produce the product of the gene of interest at high levels in a stable manner. In general, CHO cells are considered cytogenetically stable, a feature that translates to recombinant CHO cells.

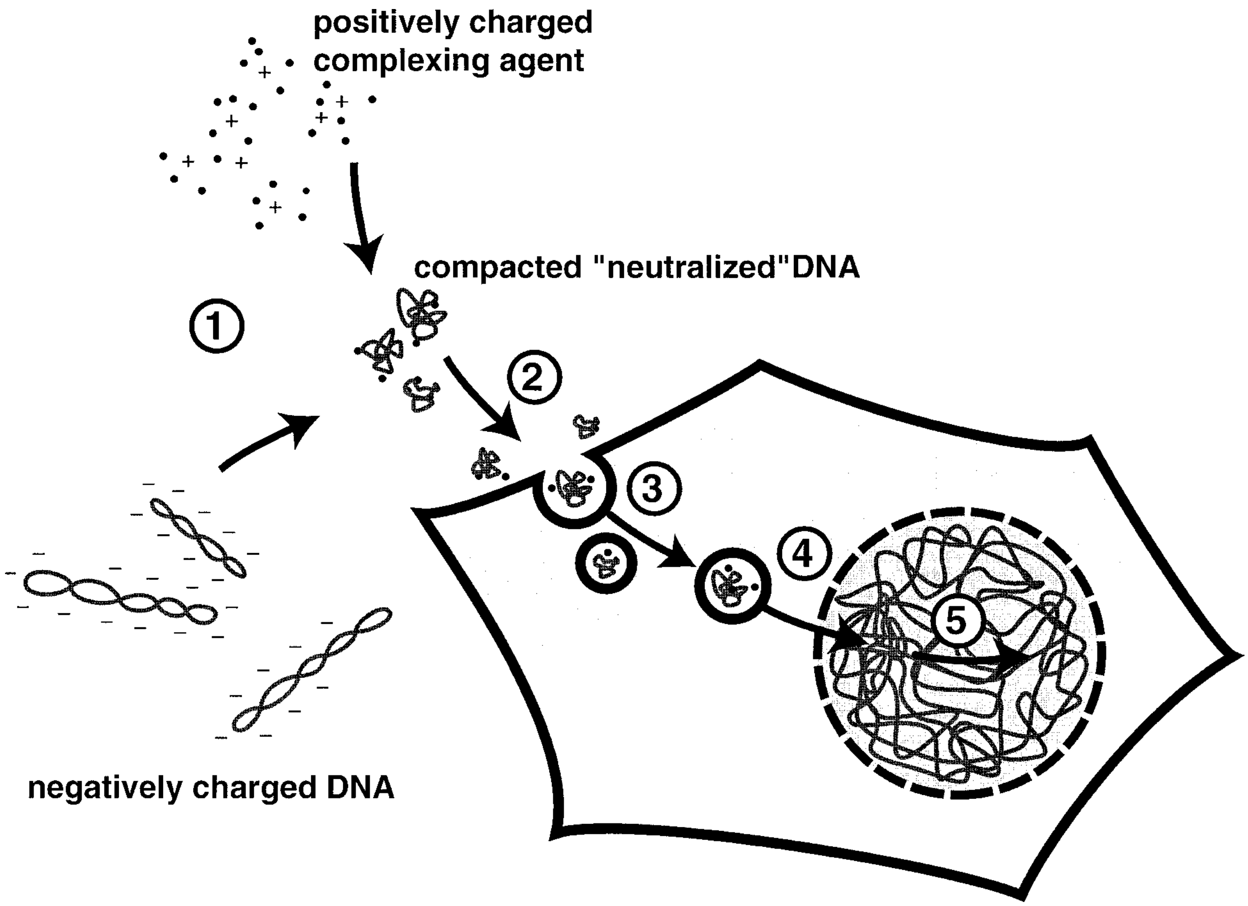


FIGURE 1 Transfection barriers: DNA for integration into chromosomes of mammalian cells has to overcome various barriers. The steps involved are numbered: (1) Neutralization and compactation of DNA, (2) association with cellular membrane and endocytotic entry of compacted DNA, (3) endosomal transfer towards nucleus and partial degradation (not shown) of DNA, (4) entry of DNA into nuclear environment, and (5) integration of DNA into chromosomal DNA.

Besides DHFR, drugs such as neomycine, hygromycine, and puromycine can be used in combination with plasmids that confer resistance to these components for the selection and identification of recombinant cells. A positive selection mechanism is based on the use of the glutamine synthetase (GS) gene in transfected plasmids when growing CHO cells in medium lacking glutamine. The GS system can be applied to other non-CHO cell lines and also allows the amplification of the transferred expression vectors in the genome of these cells.

The integration of exogenous DNA in the chromosomes of mammalian cells is poorly understood and, when using standard transfection and selection procedures, is a largely uncontrolled process. A major reason for observed variability in expression of clones isolated upon transfection is the variation in copy number and in location of integrated plasmid molecules. In spite of usually large excesses of plasmid molecules entering the cells upon transfection (10,000 to 100,000 copies), variable but small numbers of plasmid copies (1 to 100) integrate into a single chromosomal locus that is different for each clone. These two factors of variability in gene integration and probably a number of other, so far unknown phenomena make it necessary to analyze many clonal cell lines to identify a few that meet the expected expression levels. With moderate screening and nonoptimized vectors, expression levels of 0.1 to 10 pg of recombinant protein/cell/24 h can be considered acceptable for secreted proteins. However, from more stringently screened clones, including those that derive from methods that increase the copy number of integrated DNA molecules (see below), expression levels of 10 to 50 pg/cell/24 h can be obtained.

Another approach to possibly improve methods for identification of high-level expression from transgenic DNA in mammalian cells and to maintain better control over the outcome of the gene transfer experiments is targeting into defined chromosomal sites. Here, preferentially chromosomal areas of high transcription activity are desired, as are compensating or preventing trends in chromosomal silencing frequently observed when DNA

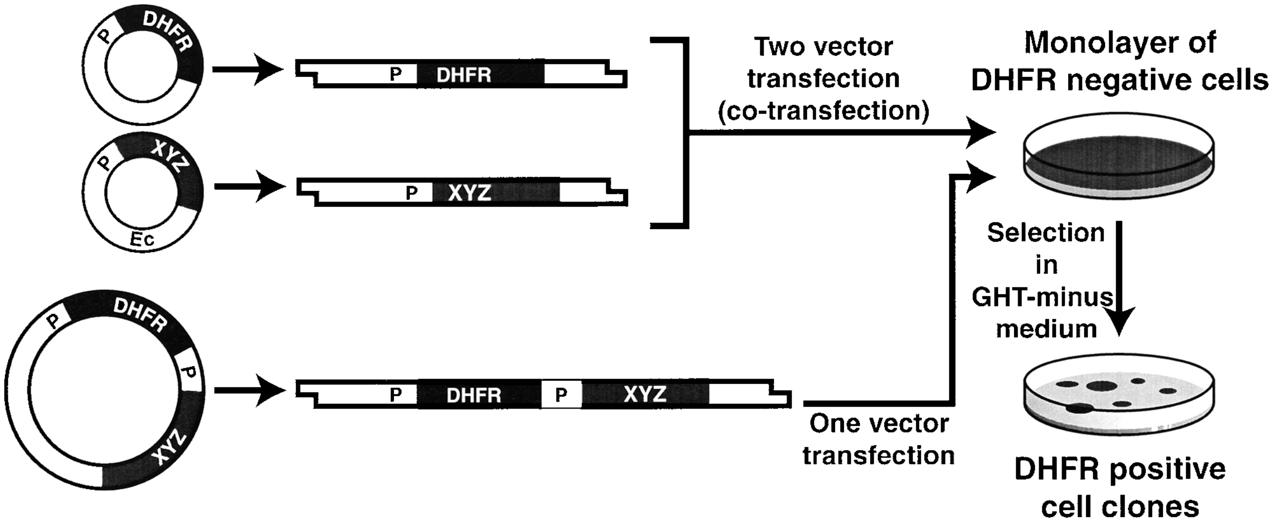


FIGURE 2 DHFR-mediated transfection: One, two, or more plasmids containing a functional DHFR expression cassette and other gene of interests can be (co-) transfected into Chinese hamster ovary cells that are DHFR negative. It is preferable to linearize the DNA to be transfected. Individual cells that have taken up and express the DHFR gene will survive and form colonies in a medium that lacks nucleotide precursor molecules [glycin, hypoxonthin, thymidin (GHT)-minus medium].

is randomly integrated in mammalian chromosomes. The targeted integration is mediated by homologous recombination whereby the targeting vector will contain sequences similar or identical to the DNA at the target site. Since targeted integration is dependent on the chance of close physical proximity of plasmid DNA and the target site, random integration will still occur at a higher frequency. To increase the frequency of targeted over non-targeted integration of foreign DNA in the mammalian genome, selection strategies can be applied in culture that employ two different selection markers—for example, a neomycin resistance gene and a herpes simplex thymidine kinase gene, both sequences adjacent to 10 to 15 kilobases of homologous DNA.

An interesting and promising approach with respect to targeting is the gene-activation method. Here, an endogenous, normally inactive gene (for example, the human erythropoeitin (EPO) gene in a human tumor cell line) is activated by the targeted integration of a strong (viral or nonviral) promoter in front of the coding sequence of the EPO gene.

3. Amplification of Transferred DNA Through Methotrexate Selection

An advantage of the DHFR/CHO system for achieving high productivity from chromosomally inserted DNA is the possibility of exposing recombinant cells to the DHFR inhibitor methotrexate (MTX). One can elevate in steps the concentration of MTX in the culture medium and isolate and subsequently establish cells that have the capacity to grow at elevated MTX concentrations. Over periods of weeks and eventually months, subclones of cell lines can be generated that are resistant to very high levels (up to 300 μM) of MTX (see Fig. 3). Such cell lines are very frequently found to have amplified segments of chromosomal regions containing the DHFR gene and the gene of interest. The result of gene amplification is not only an elevated expression of DHFR but also of the secondary gene of interest. This increase in expression rarely correlates linearly with the increase in copy number, yet 5- to 20-fold improvements of the specific productivity of stable cell lines have been found.

4. Transient DNA Transfer into Mammalian Cells for Rapid Protein Synthesis

The expression of recombinant proteins in mammalian cells from DNA sequences that have been inserted into a chromosomal site is termed *stable* expression. The generation of stable cell lines that produce at a satisfying level is usually a very time-consuming and labor-intensive process; however, it has the advantage of providing an unlimited basis—in time and scale—for product synthesis. An alternative, more rapid, and less labor-consuming approach for protein synthesis is based on *transient* gene expression. The concept of transient gene expression from mammalian cells has its molecular foundation in an efficient DNA transfer to the majority of cells in a culture, usually maintained and expanded as an adherent format in standard tissue culture flasks. All cells that have received a sufficient quantity of DNA will then begin to

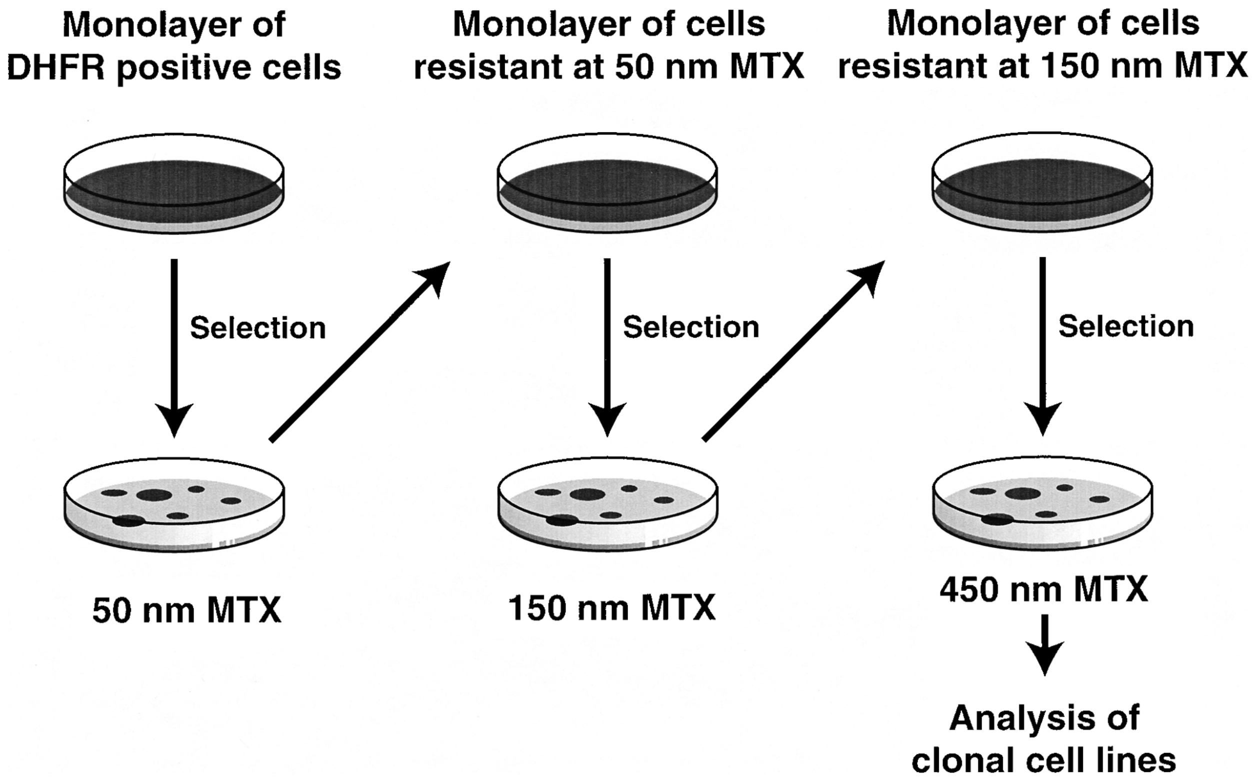


FIGURE 3 DHFR-mediated gene amplification: Cell populations containing a functional DHFR gene are exposed to stepwise elevated concentrations of methotrexate (MTX) in the culture medium for 2 to 4 weeks at each concentration. The majority of cells will die, but a number of clonal cells will multiply and form a monolayer that can then be exposed to a higher level of MTX. The majority of clones subsequently analyzed will have amplified the chromosomal DNA segments containing the DHFR gene and associated gene of interest sequences.

produce the desired protein of interest. One method for DNA transfer that has been used for this purpose for almost 30 years is the calcium phosphate DNA co-precipitation technique. Several improvements of the technique have made it very reliable and easy to use. A number of commercial transfection kits are now available, most of them based on synthetic or semisynthetic polymers, that work very well. Usually, within a few days upon transfection, the majority of cells have degraded or otherwise lost the transfected DNA. However, during this period, some DNA sequences provided by the transfection, will have been utilized by nuclear polymerases for transcription, producing mRNA and thus initiating protein synthesis. This approach is widely used in the setting of a standard research laboratory, usually with the goal to study the function of the transfected DNA within the cell. It has become a standard approach for the rapid synthesis and subsequent analysis upon purification of a desired protein. With adherent cultures, microgram quantities of recombinant protein can be produced quite easily. Efforts have been initiated more recently to perform transient gene expression at the bioreactor scale, since many studies, especially when the product is eventually intended for a pharmaceutical application, require purified protein quantities in the milligram to the hundreds of milligram range, even before clinical studies are initiated. One of the limitations in large-scale transient expression when considering a scale beyond 20 L is the quantity of DNA. With an optimized calcium phosphate method, about 1 mg of plasmid DNA is required to transfect cells in suspension cultivated at the 1-L scale. With this technique, expression levels of up to 20 mg/L of a recombinant antibody have been observed. One can expect that improvements in understanding of gene transfer and technological breakthroughs will allow production of recombinant proteins at the 100-L scale or larger in the near future.

III. THE CULTURE SYSTEM

A. Reactors for Anchorage-Dependent Cells

1. Basic Culture Units

Tissue culture flasks and tubes (glass or polystyrene plastic with special surface treatments) with surface areas of 5 to 200 cm^2 are familiar stock items of any culture

laboratory. The largest stationary flask routinely used is the Roux bottle (or disposable plastic T-flask), which has a surface area of 175 to 200 cm^2 (depending upon make and type), requires 100 to 150 mL medium, and utilizes 750 to 1000 cm^3 of storage space. Such a vessel will yield 2×10^7 diploid, or 10^8 heteroploid cells; thus, to produce a modest 10^{10} cells, over 100 replicate cultures are needed (i.e., manipulations have to be repeated 100 times) and 100 L of culture space are required. Although a few processes (mainly for classical viral vaccines) still use this methodology, the need to move to larger units is obvious and is approached by increasing the ratio of surface area to volume. The first step in scale-up usually involves a change from stationary flasks to roller bottles (i.e., a dynamic system) and is currently in widespread use for many products. Roller bottles can be up to 1750 cm^2, use 350 mL medium, and have a volume of 2.5 L (i.e., a ninefold increase in surface area, but only a threefold increase in medium and total volume). This is possible because the cells use the total internal surface area for growth. In addition, more efficient aeration occurs because the cells move in and out of the culture fluid. This method has been used industrially to produce viral vaccines, veterinary vaccines in multiples of 28,000, interferon, and EPO. It has a place in modern processes largely because it has been automated using robotic systems such as Cellmate (The Automation Partnership; Royston, Herts, U.K.). All the routine manipulations of cell seeding, media changing, bottle gassing, cell sheet rinsing, trypsinization, and cell collection by scraping can be carried out automatically, and reproducibly, with very precise volumes. Examples of products produced by this method are Varivax® (Merck varicella vaccine) and Saizen (Serono recombinant human growth hormone).

Modifications to increase the ratio of surface area to volume for the roller bottle culture include the Spira-Cell Multi-Surface Roller Bottle (Bibby Sterlin, Ltd.), extended surface area roller bottles (ESRB) (Bibby Sterlin, Ltd.), and TexturSil, a silicone rubber matrix that coats roller bottles to increase the surface area (Ashby Scientific, Ltd.; Coalville, Leics., U.K.)

The roller bottle is a well-established technique and is still widely used, in both the research laboratory and the industrial plant. Some cell lines (particularly epithelial) may not be as successfully grown in roller bottles due to streaking, clumping, or inadequate spreading over the total surface (i.e., nonlocomotory cell lines) as in stationary bottles. An alternative scale-up route is to use multisurface plate stationary systems.

2. Multisurface Plate Units

Two commercially available examples of multiple-layered polystyrene plates stacked within a polystyrene box are

1. Cell Factory (A/S NUNC; Roskilde, Denmark): The trays are 335×205 mm (600 cm^2) and can be obtained in multiples of 1, 2, 10, and 40. The largest has a surface area of 24,000 cm^2 and requires 8 L medium (equivalent to 14 large roller bottles but requiring half the incubation space and no ancillary roller equipment).
2. CellCube (Costar; Cambridge, MA, U.S.): Although this system also has parallel polystyrene trays, it is a modular closed-looped perfusion system that includes an oxygenator, pumps, and system controller (pH, O_2, level control). The unit is very compact, with the trays being only 1 mm apart, thus the smallest unit of 21,250 cm^2 is under 5 L total volume (1.25 L medium). Additional units of 42,500 cm^2 (2.5 L medium) and 85,000 cm^2 (5 L medium) are available, and four units can be run in parallel in the system, giving 340,000 cm^2 growth area.

3. High-Volume Units

These are true unit process systems, analogous in volume and performance to suspension cell fermenter vessel processes:

1. *Glass bead culture*: Packed beds of 3- to 5-mm glass spheres through which medium is continuously perfused have a demonstrated scale-up to 100 L. Spheres of 3-mm diameter pack sufficiently tightly to prevent the bed from shifting but allow sufficient medium flow up the column so that fast flow rates, which would cause shear damage, are not required. Medium can be circulated by pump or by airlift (for better oxygenation). Glass-sphere packed beds constitute a simple system that minimizes moving parts and the risk of mechanical failure and has an inexpensive and reusable substrate capable of considerable radial and reasonable vertical scale-up to operate beds of over 200-L volume using 5-mm spheres. The disadvantages are that, as spheres have the minimum surface area per unit volume, the culture will always be bulky, with most of the volume being dead space. Also, the system offers limited secreted products, as it is difficult to harvest cells from the bed, so it is ideal for long-term continuous cultures rather than batch cultures.
2. *Microcarrier culture*: Microcarriers are small particles, usually spheres 100 to 300 μm in diameter that are suspended in stirred culture medium. The technique was initiated in 1967 but required considerable developmental work to produce a range of suitable microcarriers (e.g., the Cytodex series by Pharmacia). The first industrial process based on microcarriers was for FMDV. Subsequently, a wide range of microcarriers based on gelatin, collagen, polystyrene, glass, cellulose, polyacrylamide, and silica have been manufactured to meet all situations. The key criteria in the design of effective microcarriers were to make the surface chemically and electrostatically correct

for cell attachment, spreading, and growth. The power of the method is exemplified by the following data:

a. 1 g Cytodex = 6000 cm^2 which at 2 g/L = 12,000 cm^2/L (equivalent to eight large roller bottles)

b. Scale-up to 4000 L has been achieved (=3200 roller bottles) in an environmentally controlled and optimized process.

This method opened up both industrial production opportunities and allowed research laboratories to easily produce substantial quantities of developmental products. Microcarrier culture is the most versatile, reliable, and characterized procedure for unit volume scale-up of anchorage-dependent cells. It has had widespread use for industrial processes [vaccines, interferon, tPA, and human growth hormone (hGH)] as well as many developmental uses, has been scaled-up to 4000 L, and has the potential for process intensification by perfusion with spin filters or by the use of microporous microcarriers.

B. Bioreactors for Suspension Cells

1. Laboratory Scale

The basic culture unit for suspension cells is the spinner flask, so called because it has a magnetic bar operated by standing the unit on a magnetic stirrer. Side arms are usually fitted to allow gassing with CO_2/O_2 through a filter and for sampling. The spinner flask is usually glass with a silicone- or Teflon-coated magnet and is available in sizes from 50 mL to 20 L, although 10 L should be considered the maximum practical size. It is advisable to siliconize the culture vessel and add medium supplements, such as Pluronic F-68 (polyglycol) (BASF, Wyandot) at 0.1% to protect against mechanical damage, especially if low serum concentrations are used. Stirring speeds for suspension cells are usually within the range of 75 to 250 rpm (depending upon cell type and vessel geometry), and a culture will expand from 1–2 × 10^5/mL to 1–3 × 10^6/mL in 4 to 5 days before requiring either harvesting or a medium change. Conventional spinner flasks are available from a wide range of laboratory suppliers. Modified spinner cultures include a radial stirring action (Techne), a floating impeller (Techne BR-06 Bioreactor), the Superspinner (Braun Melsungen AG), CellSpin (Tecnomara AG, Switzerland) and a dual overhead drive system to allow perfusion (Bellco). For small-scale culture (5 to 10 mL) polypropylene tubes (50 mL) placed on a horizontal circular motion shaker at 100 to 200 rpm are a useful means of carrying out process optimization studies.

2. Scale-Up

Suspension culture is the preferred method for scaling-up cell cultures as it is easier to volumetrically increase the size of a single fermenter vessel than the specialized units used for ADC. Although some cell lines will not grow in suspension [e.g., human diploid cells (HDC) lines WI-38 and MRC-5], many lines can be adapted to grow in free suspension by recognized procedures.

Several physical factors must be satisfied for successful scale-up. Good mixing is essential for homogeneity and efficient mass transfer, but during scale-up the power required to produce these conditions can cause problems. Edy effects such as the energy generated at the tip of the stirrer blade, and particularly gas bubble rupture at the media surface, are limiting factors giving rise to damaging shear forces (created by fluctuating liquid velocities in turbulent areas). As mixing efficiency increases with turbulence, a compromise has to be reached to minimize cell damage. Thus, cell cultures use large impellers running at relatively low speeds. Magnetic bars used in spinner flasks give only radical mixing, with no lift or turbulence, and scale-up means a move to marine (not turbine) impellers. Alternative systems to avoid stirring have been developed:

1. *Air-lift fermenter*: This fermenter uses the bubble column principle to both agitate and aerate a culture. Air bubbles are introduced into the bottom of a culture vessel (aspect ratio 8–12 : 1) and rise up an inner draft tube. Aerated medium has a lower density than nonaerated so the medium rises through the draft tube and circulates down the outside of the vessel (upflow and downflow of approximately equal volumes). The amount of energy required is very low and shear forces are absent; thus, it is an ideal method for fragile cells. Air-flow rates of about 300 mL/min are used. Vessels are available from 2 L upwards, with 2000-L reactors being used in industrial manufacturing processes. However, scale-up is more or less linear; a 90-L vessel requires 4 m headroom, so special manufacturing suites are needed for large-scale operations.

2. *Celligen fermenter*: This fermenter is available from New Brunswick Scientific and is designed so that the impeller acts as a fluid pump and aerator. Macrocirculation mixing is generated by the hollow central shaft filled with three rotating horizontal jet tubes in a low shear bulk movement of cells and medium (cell-lift effect).

The other principal physical factor to be accommodated is aeration. Sparging can cause cellular damage unless carried out at very low rates. General guidelines are to use large bubbles (1 to 3 mm in diameter) and low flow rates (5–10 cm^3/min) which means relatively inefficient mass transfer, especially at the low stirring rates used. Even though cell bioreactors have an aspect ratio of less than 2:1 in order to maximize surface diffusion, supplying sufficient oxygen during scale-up is a limiting factor. Strategies to address this have included a surface aerator,

multiple nozzle injection on the medium surface, caged aerator, spin exchange aerator, external and internal loop oxygenator, and membrane tubing arranged in the vessel or stirring system. The use of closed perfusion loops through external reservoirs is also an efficient means of oxygenation as long as the perfusion rate is fast enough, which requires a very efficient spin filter or other cell separation device in the culture.

3. Large-Scale Bioreactors

Large-scale bioreactors are based on either the fermenter tank or the airlift principle. Stirred tanks have been operated at 8 to 10,000 L for the production of interferon, tPA, recombinant antibodies, and DNase, and airlift at 2000 L principally for monoclonal antibodies. The preference is for tank fermenters, as they are a well-tried and familiar production method using conventional production plant facilities and have been refined over the past 30 years; however, most processes run at scales between 50 and 500 L.

Containment of the process to prevent ingress of contaminating factors and the release of materials that may affect the process worker and environment is a key factor in these processes. Microbiological containment of small-scale systems (below 20 L) in class III microbiological cabinets is possible, and a larger scale fermentation plant can be contained to P3 standards at least up to the 150-L scale.

The engineering complexity and considerable resource investment in money and time to get a large-scale production process built and operating mean that many new products are still brought to market using replicate small-scale systems. This allows a quicker market entry, but it does involve a complete re-registration of the product if, at a later stage, the product is moved to a large-scale production process.

C. High-Cell-Density Bioreactors

In this section, the processes that support cells at significantly greater densities than the classical $2–3 \times 10^6$/mL level are described. In the human body, cells in tissues are found in the order of $2–3 \times 10^9$ cells/cm^3, and this proportion has always been a target for *in vitro* systems. However, 2×10^8 cells/mL have so far been the practical limit of density scale-up, but even so this is a 100-fold increase. To achieve higher unit cell density requires (1) perfusion to supply nutrients and remove waste products, and (2) cell immobilization in order to perfuse media at a fast enough rate without damaging or washing the cells out of the culture. A huge range of devices have been developed to meet these requirements, of which hollow-fiber bioreactors, spin-filter perfusion, and microporous microcarriers are the most successful in terms of acceptance and use.

1. Hollow-Fiber Bioreactors

This method was pioneered by Knazek in 1972 using ultrafiltration capillary fibers. They can be considered analogous to a blood vascular system as the fibers selectively (by molecular weight cut-off) allow passage of macromolecules through the spongy fiber wall (60 μm) as the medium flows continuously through the lumen (200-μm diameter). Cells are kept in the extra capillary space. A unit consists of thousands of fibers "potted" at either end in a cylindrical housing and capable of supporting both ADC and suspension cells at $1–2 \times 10^8$/mL. This concept has been widely developed with many commercial units available and has been used particularly for producing monoclonal antibodies. The principal limiting factor in their exploitation is the difficulty in scaling-up beyond very small volumes (25 mL). Thus, this method gives a very large surface-to-volume ratio, allows continuous removal of waste products and supply of nutrients, supports high cell densities with a tissue-like architecture, and allows a concentrated product to be harvested. The disadvantages are diffusional limitations causing culture inhomogeneity and cell necrosis, process control complexity, and difficulty in sterilization.

2. Spin Filters

As all static filters within a culture become blocked, sooner rather than later, the development of a rotating filter that creates a boundary effect, thus delaying cell attachment and filter clogging, has been an important step forward. These filters provide a simple technical solution to scaling-up process intensity within well-established stirred fermenters. They are limited in that clogging does eventually occur and in the perfusion rate that can be maintained. However they are particularly useful for microcarriers, as a far larger mesh can be used, thus reducing the onset of filter clogging and permitting cell densities of 2×10^7/mL to be attained.

3. Microporous Microcarriers

Microcarrier culture has proven to be the most effective scale-up method for ADC, despite its limitations (critical procedures, low surface area-to-volume ratio of a sphere). To increase the surface area porous particles were developed. The initial particle was the Verax microsphere which was 500 μm in diameter and manufactured from bovine collagen. The interconnecting channels of 20- to 40-μm diameter provided an internal open volume of 80% of the sphere. The spheres were fluidized at 75 cm/min upward flow, and cell densities in excess of 10^8/mL intrasphere volume were achieved (equivalent to 4×10^7/mL in the bioreactor). The sphere matrix provided a huge surface

area for attachment of ADC, but was equally suitable for suspension cells entrapped in the pores. The system was scaled up to 24 L routinely and to 200 L in the company's production unit. It was used for over 85 different cell lines producing many cell products (including tPA, proUK, EPO, IFN, IL, factor VIII, various immunoglobulins).

Many microporous beads are available, as are systems such as the Pharmacia Cytopilot using Cytoline porous microcarriers. There is a preference for microporous microcarriers that could be used in stirred, rather than fluidized, bioreactors, and these are now available (Cellsnow and ImmobaSil). ImmobaSil is of particular value as it is extremely permeable to oxygen, is a non-animal product safe from bovine contaminants, is robust, and can be used in all culture modes.

An alternative to using stirred/fluidized microporous carriers is to use a fixed bed of porous glass spheres (e.g., Siran). Fixed bed reactors of solid glass spheres have many advantages, and the disadvantage of low cell density can be overcome by using 5-mm Siran spheres. These have a surface area of 75 m^2/L with interconnecting pores and channels of 60 to 300 μm. They can be stacked in 5-L beds and perfused at 5 linear cm/min to provide a high feed rate without washing out or damaging the cells (protected from shear within the particles). The versatility and usefulness of this technique have been demonstrated for a range of suspension and anchorage-dependent cells, resulting in a tenfold higher productivity over equivalent systems.

A characteristic of all microporous carrier systems is that cell-specific productivity is always higher, presumably due to the favorable environment of cells packed together in almost tissue-like density. Commercial fixed-bed reactors are available (Meredos GmbH, D-37120 Bovenden).

The advantages of porous carrier culture are summarized in Table III. This technology is truly universal as it is equally suitable for suspension and ADC, is flexible in the range of bioreactor systems it can be used in, is the only truly scalable high-cell-density system available, and provides a micro-environment that stimulates cell product expression.

TABLE III Advantage of Microporous Microcarriers

1. Unit cell density 50- to 100-fold higher than free suspension
2. Suitable for both attached and suspension cells
3. Can be used in fluidized and fixed-bed reactors and in stirred suspension
4. Elimination of seed chain steps by *in situ* 100- to 250- fold seed expansion
5. Efficient diffusion into a sphere (30% diameter penetration = 70% of volume)
6. Protection of cells from shear
7. Three-to fivefold increase in specific cell productivity
8. Easily derivitized three-dimensional structure for specialized cells
9. Capable of long term (>100 d) culture with continuous harvesting
10. Scale-up potential as compared with analogous systems (microcarrier to 4000 L)

IV. THE CULTURE PRODUCTION PROCESS

The basic principles common to most processes are summarized in Fig. 4. The components are

1. *Seed banks*: Both the cell line and virus for vaccines have to be laid down in a fully tested and characterized bank. Thus, each production batch will be initiated from identical cells (and virus) known to be viable and contamination free.
2. *Cell seed expansion*: A series of culture steps is needed to expand the cell seed ampoule (e.g., 5 million cells) to production size (range 10^9 to 10^{13} cells). For HDC, this is accomplished in steps of a split level of 1:2 or more usually 1:4 through a series of flasks, roller bottles, and possibly cell factories (A/S Nunc). Other cell types are split at a 1:5 to 1:20 ratio. A similar build-up is needed for the virus seed.
3. *Production*: The production culture may be a batch of several hundred roller bottles, 30 to 50 cell factories, or a single bioreactor for suspension (100 to 10,000 L) or microcarrier (50 to 500 L) cells. Although batch-type production is still the most common process, continuous processes where the product is harvested daily over a long period (20 to 100 d) are being increasingly used. Culture systems based on hollow fibers, porous microcarriers, or other immobilization techniques are used for continuous perfusion processes. During the production phase, the virus seed or a promoter (e.g., for interferon) may be added.
4. *Harvesting*: If the product is intracellular, then the cells have to be harvested (trypsin and/or EDTA), washed, and concentrated by centrifugation. Extracellular (secreted) products just require the collection of the culture supernatant.
5. *Downstream processing*: Intracellular products have to be extracted from the cells (by sonication, freeze thawing, and/or homogenization), and separated from the cells (centrifugation or filtration). Extracellular products require concentration and separation from the bulk supernatant.
6. *Formulation*: The product is added to a medium with protective and stabilizing agents and then usually freeze-dried.
7. *Quality control*: Throughout the process, prescribed samples are taken for a range of quality control tests to show safety, efficacy, and consistency in the process and product.

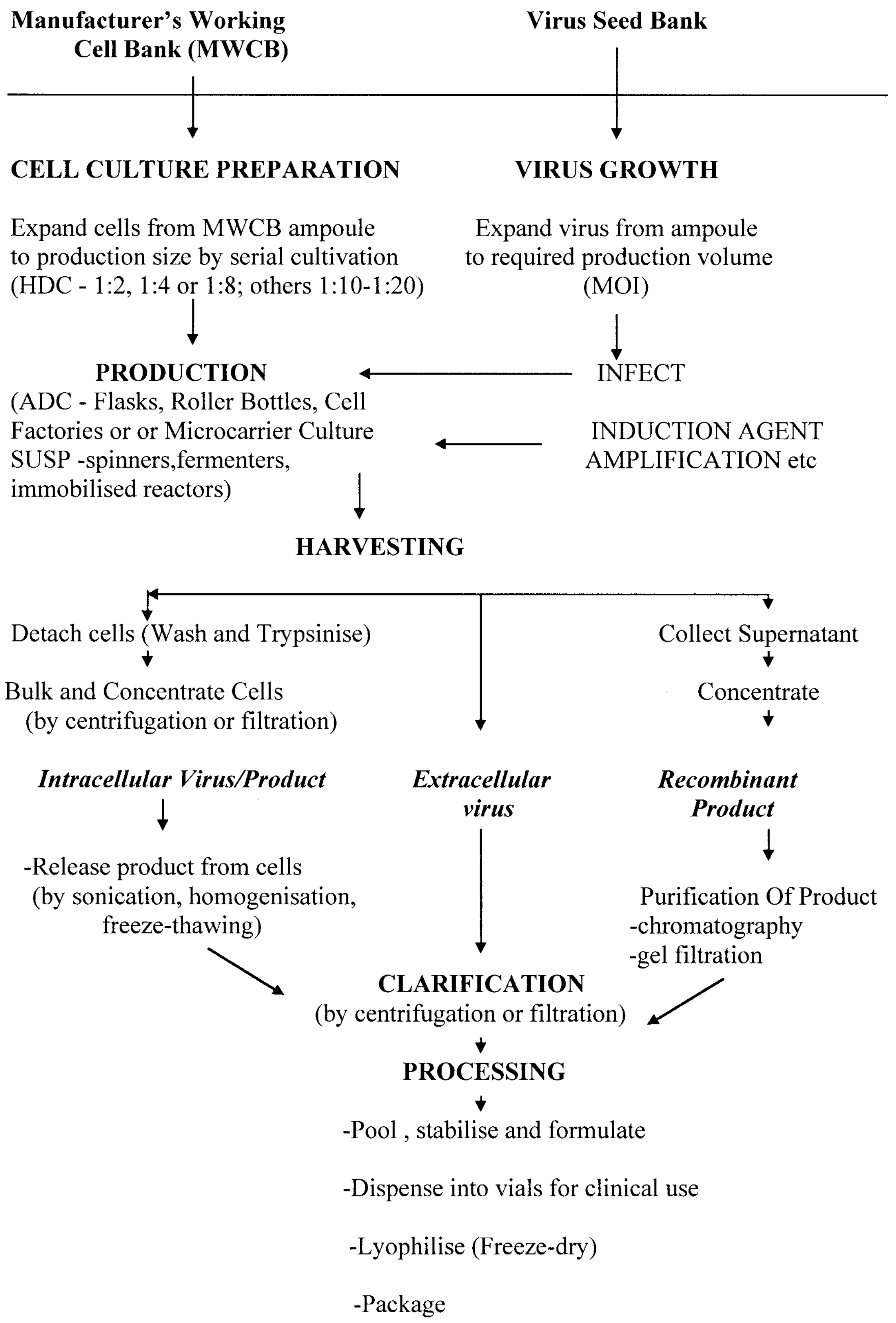


FIGURE 4 Production process schematic for mammalian cells (vaccines, antibodies and recombinant products, etc.).

The use of mammalian host systems for the production of pharmaceutically relevant products has been discussed, at times quite controversially, by regulatory control agencies, academic institutions, and private companies. In all these discussions, the transfer to patients of agents (globally referred to as "adventitious" agents) has been the reason for concern. The three main types of agents are viruses, DNA-containing oncogenes, and any other pathogenic activity and contaminating proteins that may give rise to immune reactions or any other adverse reaction. In fact, some of the early experiences, especially with primary cells, resulted in the transfer of viruses of known or unknown origin to human patients (e.g., SV40 virus in polio vaccines). The sophistication of modern purification techniques and the accuracy of assay measurements at the picogram level allow substates to be used that were considered a hazard 15 years ago, as quality control emphasis is now on the purity of the final product.

V. CELL PRODUCTS

A. Viral Vaccines

The first cell-based vaccine was for polio and was produced in monkey kidney cells. A series of HDC vaccines were licenced during the 1960s. The first recombinant vaccine was against hepatitis B and currently the primary target is human immunodeficiency virus (HIV). Other viral diseases with trial vaccines are hetres simplex virus (HSV), rous sorcoma virus (RSV), cytomegalovirus (CMV), influenza, and rotavirus. The vaccine field has expanded from the area of infectious diseases to include cancer (particularly melanoma, also breast, colorectal, ovarian, and B-cell cancers), rheumatoid arthritis, multiple sclerosis, and contraception. The vaccine itself is showing a dynamic evolution from the original whole virus (dead or attenuated), through subunits (native and recombinant), genetically deleted viruses, to the future aim of DNA vaccines and mucosal immunity.

B. Antibodies

The production of monoclonal antibodies, first at the research level and then for diagnostics (including *in vivo* imaging) from the mid-1980s gave a huge impetus to industrial animal cell biotechnology. The use of monoclonal antibodies expanded from small-requirement (dose) diagnostics to large-dose therapeutics for HIV, cancer, allergic diseases, arthritis, renal prophylaxis, septic shock, transplantation, asthma, CMV, and anti-idiotype vaccines. The development of recombinant monoclonal antibodies was largely driven by the need to "humanize" the product to prevent immunological incompatibilities that lead to very short half-lives and make only single-dose treatment possible. The field has moved on to the use of adoptive immunotherapy, where the patient's cells are altered and grown *in vitro* and perfused back into the patient. Many novel products have been developed—for example, the CD-4 receptor, which is a combination of the genes coding for the soluble form of the CD-4 receptor with the gene sequence for IgG molecules which results in a soluble receptor for HIV.

C. Immunoregulators

The production of alpha-interferon in Namalwa cells by Wellcome was the first licenced use of a cancer cell substrate for a human biological. The unrestricted growth in suspension culture of these cells allowed the first multithousand-Liter (8000 L) unit process to be developed for human products. The knowledge gained from producing FMDV vaccine in suspension BHK cells was of great benefit in developing this process. A wide range of both interferons and interleukins occurs naturally, and to date both alpha- and gamma-interferons and interleukins 2, 3, 4, 6, 11, and 12 have been manufactured in culture.

D. Recombinant Products

The fact that cells with specialized *in vivo* functions, such as endocrine cells secreting hormones, could not be grown and replicated in culture with retention of their specialized properties has always been a great disappointment, not only for advancing medical studies but also for using cells to manufacture naturally occurring biologicals. Thus, genetic engineering techniques that allow the gene(s) responsible for production of required biologicals in a highly differentiated (nonculturable) cell to be inserted into a fast-growing robust cell line have opened up numerous possibilities for exploitation by animal cell technology (Table I).

The emerging pharmaceutical biotech-industry realized in the mid-1980s that the expression capacity of microbial systems, and in particular of *Escherichia coli*, was limited when the protein of interest required a complex folding with multiple disulfide bridges and secondary modifications such as glycosilation, fucosilation, or other post-translational processing. Simply, the majority of somewhat larger proteins could not be expressed in a functional and structurally correct form in microbial systems. Human tPA, a protein with 528 amino acids, 15 disulfide bridges, and three glycosilation sites is in this category. Initial attempts to express functional tPA in *E. coli* failed, while efforts to produce moderate quantities of this protein in CHO cells were successful. The naturally adherent CHO cells were scaled up for an industrial robot-driven production

process in roller bottles, an approach that is useful due to its simplicity and reliability. Such an approach is still being used today for a number of processes, including one that was developed later for another highly glycosilated protein, the red blood cell growth factor erythropoeitin (EPO). EPO is a hormone produced by the kidney that controls the maturation of red blood (erythroid) cells and has clinical applications in anemia due to chronic renal failure. The product was licenced in 1989 (by Amgen) as Epogen and in 1990 as Epogin.

The productivity of roller bottle processes is usually limited, since the only controlled environmental factor is the temperature. For human recombinant tPA made in CHO cells through a roller bottle process, expression levels could not be elevated beyond 5 to 10 mg/L of cell culture supernatant, revealing a major supply restriction for the anticipated global markets for the novel thrombolytic agent. This resulted in the successful efforts to adapt CHO cells to growth in suspension as single cells, thus allowing the use of well-established bioreactor stirred-tank technology, overcoming both a volumetric limitation and titer limitation.

The first process to be developed for a recombinant pharmaceutical from a mammalian host was finally produced at the 10,000-L scale with a product titer in the medium of about 50 mg/L and was licenced as Activase/Actilyse in 1987. To put this in perspective, nonrecombinant endothelial cells and other *in vivo* rich sources (e.g., human uterus) contain only 1 mg tPA per 5 kg uterus (0.01 mg purified tPA per uterus). Some tumor cell lines such as Bowes melanoma secrete tPA at a higher rate (0.1 mg/L), but this amount still was not economical for a production process and (at that time) was considered unsafe coming from a human melanoma. tPA is a product necessary for dissolving blood clots and is much needed for the treatment of myocardial infarcation and thrombolytic occlusions. Alternative products, urokinase and streptokinase, were less specific and could cause general internal bleeding and other side effects. tPA was therefore an ideal model product for cell technology and an example of a high-activity/low-concentration product that was clinically in demand. Genetic engineering not only allowed the product to be produced in a relatively safe cell line but was used to amplify cell production (50 mg/10^9 CHO cells/day) from the low native secretion rates. By the year 2000, more than 20 large-scale production processes had been developed, based on the model of the rtPA process, utilizing the DHFR/gene-amplification/suspension/stirred-tank bioreactor technology. Recently, reports have revealed expression levels of 2 g/L and higher for recombinant human antibodies produced by suspension-adapted CHO cell lines.

E. Cell and Tissue Therapy

1. Cell Therapy

Cell therapy is the replacement, repair, or enhancement of biological function of damaged tissue or organs achieved by transplantation of cells to a target organ by injection (e.g., fetal cells into the brain of patients with Parkinson's or Alzheimer's disease), or by implantation of cells selected or engineered to secrete missing gene products. The first recorded application was growing keratinocytes from a small skin biopsy into large cell sheets, which were then grafted onto burn patients. This application has now advanced to commercial production of dermal replacement products such as Dermagraft. To avoid destruction of implants by the host's immune system, encapsulation of the transplant cells in semipermeable devices is widely used. Examples include pancreatic islet cells for diabetes, chromaffin cells for chronic pain, and genetically engineered BHK cells secreting neurotrophic factors for neurodegenerative diseases. It has not yet been possible to replace the liver or kidney, but artificial organs situated outside the patient containing primary or recombinant cells through which the patient's blood is perfused have been developed. Dialysis techniques only remove the toxic products, whereas the cells in the artificial organs perform biotransformations—as well as degrading toxic products, they additionally regenerate many essential metabolites which are returned to the body.

The future of cell therapy is expected to be based on stem cells (self-renewing cells that give rise to phenotypically and genotypically identical daughter cells). Stem cells develop via a "committed progenitor stage" to a terminally differentiated cell. They are multipotent—that is, they are able to develop into a wide range of tissues and organs, but only fertilized germ cells are totipotent—able to give rise to all cell tissues in the body. Control of the development of stem cells into the required tissue or to stimulate quiescent "committed progenitor cells" of the required tissue with the relevant growth factors and hormones would allow the most effective cell therapy possible. This approach is causing some ethical controversy, as the most suitable source of stem cells is cloning them from human embryos. The technique is to extract the genetic material from an adult patient needing transplantation, introduce it into a human egg with its nucleus removed, and grow the embryo *in vitro* for eight divisions until stem cells can be treated with growth factors to form the required tissue (e.g., pancreas, nerve, etc.). Interest in replacing organs or damaged tissue with the help of cells that have been cultivated is expanding rapidly due to the finding that the nucleus of adult, fully differentiated mammalian cells can, under conditions, be reprogrammed to allow even the cloning of an animal.

2. Gene Therapy

Gene therapy has the potential for treating a very wide range of human diseases but since the first somatic gene therapy product [T-lymphocyte-directed gene therapy of adenosine deaminase deficiency (ADA-SCID)] went into trial in 1990 progress has been disappointingly slow, although there are over 300 clinical products at some stage of clinical trial. One problem is the development of safe and efficient gene-delivery systems, which require careful regulation for safety. This effort has largely concentrated on engineering viruses as vectors of therapeutic genes

Three modes of gene delivery are possible:

1. *Ex vivo*: Removal of the cells from the body, incubation with a vector, and return of the engineered cells to the body (mainly applicable to blood cells),
2. *In situ*: Vector is placed directly into the affected tissues (e.g., infusion of adenoviral vectors into trachea and bronchi for cystic fibrosis patients).
3. *In vivo*: Vector is injected directly into the blood stream (this method is a goal but has not yet been used clinically).

The target diseases currently undergoing clinical trial are

1. Cancer (melanoma, colon, renal cell, neuroblastoma, ovarian, breast, lung)
2. Genetic diseases (cystic fibrosis, Gaucher's disease, SCID)
3. Viral (acquired immune deficiency syndrome, AIDS)

F. Other Products

A product area of increasing interest and potential is the cell-adhesion molecules (CAMs). These are molecules that mediate cell–cell and cell–matrix interactions and are being developed as drugs against inflammatory diseases. They also have the potential to treat metatastic diseases, atherosclerosis, and microbial infection. Chemokines present at sites of inflammation and disease bind leukocyte receptors and activate a family of CAMs known as integrins. Sequential activation and interaction of multiple CAMs in the inflammatory process offer many targets for drug intercession. Target products include antisense and antagonists. Examples of such drugs undergoing trial are Cylexin (reperfusion injury), Integretin (arterial thrombosis, angina), and Celadin (inflammatory diseases). There are over 30 companies developing CAMs which are in various stages of preclinical and clinical trial.

An important and topical new application is gene targeting into primary mammalian cells for the generation of transgenic animals. Mouse embryos have been found to provide a type of cell known as embryonic stem cells. These cells can be cultivated, genetically modified in culture, and subsequently inserted into a developing embryo. These modified cells can contribute to all of the cells of the mouse embryo, including germ cells, providing the opportunity to transfer a genetic character into a strain of laboratory mice. Recently, based on the breakthrough discovery of Wilmut and Campbell when cloning the sheep "Dolly" from an adult mammary tissue cell, gene constructs of interest have been inserted into cultured fetal fibroblasts of sheep that were subsequently inserted into enucleated eggs. Using a targeting vector and selection with a neomycin selection marker, the gene of interest for human alpha-1-antitrypsin (AAT) has been inserted into the chromosomal locus of a pro-collagen gene. Of 16 transgenic fetuses and lambs analyzed, 15 were confirmed to contain the transgene. One individual produced 650 mg/L of AAT in the milk from a single inserted copy of the AAT gene. It is to be expected that these techniques for gene transfer into mammalian cells will be expanded for the commercial use of transgenic animals for the production of desirable proteins, as well as for the generation of more useful human disease models in mammals larger than mice.

VI. APPLICATIONS OF CELL CULTURE

In addition to being used as a substrate for the manufacture of pharmaceutical products, cells are used in a wide range of other applications including diagnostic virology, aging, cell physiology and metabolism, immunology, cell genetics, and cancer. Testing in the fields of toxicology and pharmacology is also important to lead to more controlled experiments and reduce significantly the need for using animals. Cell cytotoxicity assays are used to (1) identify potentially active therapeutic compounds, (2) to identify the mechanism by which the compound exerts its toxic effect, (3) to predict anticancer activity, (4) to identify possible target cell populations, (5) to identify the toxic concentration range, and (6) to investigate the relationship of concentration to exposure time.

VII. CONCLUSION

Mammalian cell culture is used widely, from small-scale applications in the laboratory that enable research into cancer and many other medical conditions, through screening and assay procedures for identifying new drugs and testing the safety of existing products, as a basis for tissue replacement and gene deficiency correction, to a production of up

to 10,000-L batches of pharmaceuticals. The development of media containing specific growth factors, together with the sophistication of genetic techniques and implantable materials, has allowed cell culture to evolve to the use of the cell itself as a product (or means of producing a substrate) for use in biological processes.

SEE ALSO THE FOLLOWING ARTICLES

Biomaterials, Synthesis, Fabrication, and Applications • Bioreactors • Cell Death (Apoptosis) • DNA Testing in Forensic Science • Gene Expression, Regulation of • Hybridomas, Genetic Engineering of • Immunology—Autoimmunity • Metabolic Engineering • Tissue Engineering

BIBLIOGRAPHY

Campbell, K. H. S., McWhir, J., Ritchie, W. A., and Wilmut, I. (1996). "Sheep cloned by nuclear transfer from a cultured cell line," *Nature* **380,** 64–66.

Doyle, A., and Griffiths, J. B. (1998). "Cell and Tissue Culture: Laboratory Procedures in Biotechnology," John Wiley & Sons, New York.

Doyle, A., and Griffiths, J. B. (2000). "Cell and Tissue Culture for Medical Research," John Wiley & Sons, New York.

Freshney, R. I. (1994). "Culture of Animal Cells," Wiley-Liss, New York.

Hunkeler, D., Prokop, A., Cherrington, A., Rajotte, R., and Sefton, M., eds. (1999). "Bioartificial organs. II. Technology, Medicine and Materials," *Annals of the New York Academy of Sciences* **875**.

Jenkins, N., ed. (1999). "Animal Cell Biotechnology," Humana Press, Clifton, NJ.

Masters, J. R. W. (2000). "Animal Cell Culture—A Practical Approach," Oxford University Press, London.

Shepherd, P., and Dean, C. (2000). "Monoclonal Antibodies—A Practical Approach," Oxford University Press, London.

Spier, R. E. (1999). "Encyclopedia of Cell Technology," John Wiley & Sons, New York.

Wurm, F. M. (1997). "Aspects of Gene Transfer and Gene Amplification in Recombinant Mammalian Cells." *In* "Genetic Manipulation of Mammalian Cells" (Hj. Hauser and R. Wagner, eds.), pp. 87–120, Walter de Gruyter, Berlin.

Wurm, F. M. (1999). "Chinese hamster ovary cells, recombinant protein production." *In* "The Encyclopedia of Bioprocess Technology: Fermentation, Biocatalysis and Bioseparation" (D. Flickinger, ed.), pp. 570–581, Wiley, New York.

Wurm, F. M., and Bernard, A. (1999). "Large-scale transient expression in mammalian cells for recombinant protein production." *Current Opinion in Biotechnology* **10,** 156–159.

Metabolic Engineering

Jens Nielsen
Technical University of Denmark

I. Background
II. Introduction
III. Molecular Biology Tools
IV. Metabolic Network Analysis
V. Metabolic Control Analysis
VI. Tools from Functional Genomics
VII. Applications of Metabolic Engineering
VIII. Future Directions

GLOSSARY

Bioinformatics The use of informatics to upgrade biological information, e.g., to identify genes in sequenced genomes, to predict protein structures, and to construct whole cell models.

DNA arrays Glass slides or other surfaces where oligonucleotides or cDNA are fixed in very small spots. The surface is organized in an array where each element in the array represents a specific oligonucleotide sequence. DNA arrays (or DNA chips) can be used to quantify gene expression of many genes in parallel—in some cases the expression of all genes within a genome.

Functional genomics The science that aims at identifying the function of genes with unknown function, often referred to as orphan genes.

Heterologous gene A gene that is not naturally present in the host, e.g., a human gene inserted in a bacteria is a heterologous gene.

Metabolic flux analysis Analysis technique that enables quantification of the fluxes through the different branches of the metabolic network within a given cell.

Metabolic network The network of biochemical reactions functioning within a given cell.

Proteome The total pool of different proteins present within a given cell.

Transformation vector A circular or linear piece of DNA that can be used to insert a gene within a chromosome or as a self-replicating piece of DNA within a given cell.

METABOLIC ENGINEERING is the introduction of directed genetic changes with the aim of improving microbial, plant, or animal cells for the production of various products. Besides genetic engineering, it involves detailed physiological characterization of the cells, and many new experimental techniques have been developed in connection with metabolic engineering. These techniques aim at obtaining a sufficiently detailed characterization of the cellular function to identify targets for genetic improvements.

Among the techniques applied are metabolic flux analysis, metabolic control analysis, DNA arrays, proteome analysis, and metabolite profiling.

I. BACKGROUND

Fermentation based on microorganisms, plant cells, animal cells, and mammalian cells is currently used to produce a wide variety of products, ranging from bulk chemicals and materials to highly valuable pharmaceuticals. Traditionally, optimization of fermentation processes involved improved design of the bioreactor and introduction of advanced schemes for controlling the process. Parallel to this work, the properties of the organism applied in the process was improved through random mutagenesis followed by screening for better strains. This is especially demonstrated with the domestication of the baking and brewing strains of *Saccharomyces cerevisiae*. Here novel strains have been obtained through classical methods of mutagenesis, sexual hybridization, and genetic recombination. Also, in the development of the penicillin production several rounds of mutagenesis and screening have resulted in strains with improved yield of the secondary metabolite, and Fig. 1 illustrates a typical development in the performance of an industrial strain lineage of *Penicillium chrysogenum* applied for penicillin production.

II. INTRODUCTION

The first successful genetic engineering of *Escherichia coli* by Cohen, Boyer, and coworkers in 1973 paved the way for a completely new approach to optimization of existing biotech processes and development of completely new ones. Shortly after were implemented several industrial processes for production of recombinant proteins, e.g., the production of human insulin by a recombinant *E. coli*. With the further development in genetic engineering techniques, the possibility to apply this for optimization of classical fermentation processes soon became obvious, and through introduction of directed genetic modifications by rDNA technology this has enabled a far more rational approach to strain improvement than the classical approach of mutagenesis and screening. In 1991 this led Bailey to discuss the emerging of a new science called metabolic engineering, which he defined as "the improvement of cellular activities by manipulations of enzymatic, transport, and regulatory functions of the cell with the use of recombinant DNA technology." Initially metabolic engineering was simply the technological manifestation of applied molecular biology, but with the rapid development in new analytical techniques and in cloning techniques, it has become possible to rapidly introduce directed genetic changes and subsequently analyze the consequences of the introduced changes at the cellular level. Often the analysis will point toward an additional genetic change that may be required to further improve the cellular performance, and metabolic engineering therefore involves a close integration between analysis of the cellular function and genetic engineering as illustrated in Fig. 2.

According to the cycle of metabolic engineering, there is a continuous improvement of the cellular properties through several rounds of genetic engineering. Depending on the process and aim, one may start at different locations in this cycle. Thus, for production of a heterologous protein, for production of a new metabolite by pathway extension, or for extension of the substrate range for the applied microorganism, it is always necessary to start with the synthesis step. However, if the aim is to improve the yield or productivity in an existing process, it is necessary first to analyze the pathway involved in forming the product, and how this pathway interacts with the overall cell function, i.e., one should start with the analysis step. It is always desirable to optimize the yield or productivity in industrial processes, and the analysis step therefore always plays a very prominent role—also in those cases where the first step is to construct a recombinant strain that produces the product of interest. It is clear that by passing through the cycle of metabolic engineering one gains a significant insight into cellular function, and this is one of the reasons that metabolic engineering today interact closely with the discipline of functional genomics, where the aim is to assign function to orphan genes in completely sequenced genomes (see also discussion later). However, since metabolic engineering requires availability of the proper tools for genetic modifications and for analysis of cellular function, developments in genomics and analytics have been one of the main reasons for the rapid expansion of the field of metabolic engineering in

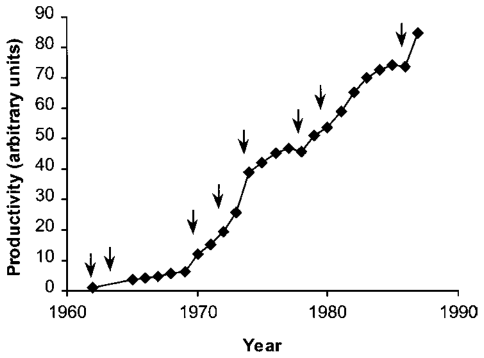


FIGURE 1 Increase in productivity (output rate/unit volume, arbitrary units) of penicillin G production by Gist-brocades, Delft (now DSM), in the period between 1962 and 1987. The introduction of new strains is marked with arrows.

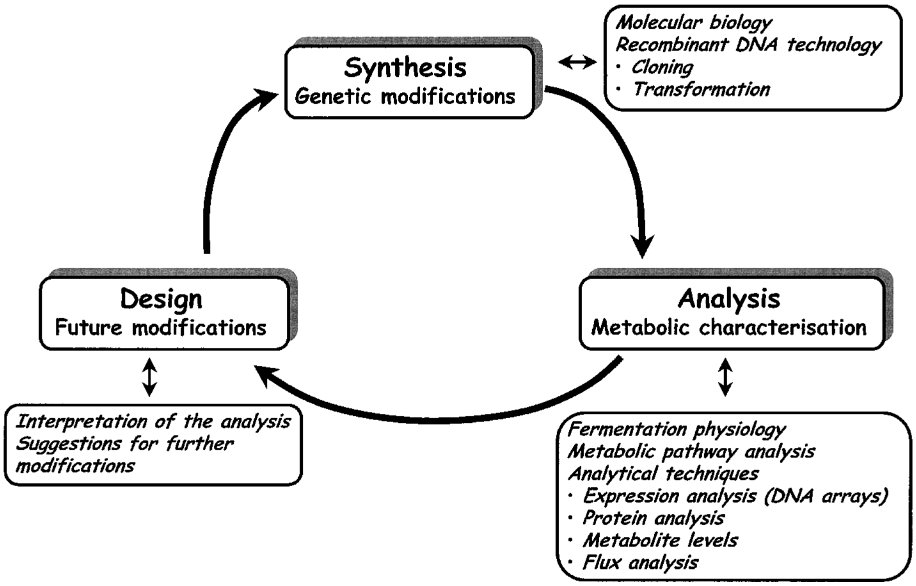


FIGURE 2 The cycle of metabolic engineering.

recent years. Thus, genetic engineering techniques have been facilitated and the sequencing of several complete genomes and developments in bioinformatics has speeded up the process of gene cloning and transformation. Today many different microorganisms have been completely sequenced—including several industrially important microorganisms like *Bacillus subtilis, E. coli*, and *S. cerevisiae*, and the list is rapidly growing (for a *up to date* list, see www.genome.ad.jp). Table I compiles some of the tools that are often recruited in metabolic engineering, and in the following some of these tools are discussed in further details.

III. MOLECULAR BIOLOGY TOOLS

For introduction of genetic modifications it is pivotal to have suitable strains and vectors that enable rapid transformation with a high transformation efficiency. For a few microorganisms, e.g., *S. cerevisiae* and *E. coli*, there are efficient transformation vectors available that enable either rapid chromosal integration of DNA or expression of genes through the use of high copy number plasmids. Furthermore, for these organisms there are many auxotrophic strains available, which facilitates the genetic engineering significantly. For *S. cerevisiae* there is a very high degree of homologous recombination, and this enables high frequency of directed DNA integration into the chromosome. For *E. coli* many special tools have also been developed, e.g., segregationally stable plasmids present in low copy numbers that enables rapid cloning of different genes and methods for stabilization of mRNA of heterologous genes through introduction of specific hairpins. For other microorganisms the necessary tools may be available, but the transformation efficiency is often low, and the genetic engineering is much more cumbersome. For many industrially important organisms, e.g., *P. chrysogenum*, *Aspergillus* species, and *Corynebacterium glutamicum*, there are also suitable transformation vectors, and even though there are less auxotrophic strains available (at least in industrial strain backgrounds) there are typically several dominant markers available.

TABLE I Tools of Metabolic Engineering

Molecular biological tools	Analytical tools
Methods for rapid gene cloning	Gene expression analysis (DNA arrays, Northern's)
Stable plasmids	Protein level analysis (2D gels, Western's)
Different promoters (inducible, strong, etc.)	Metabolite profiling
Site-directed chromosomal integration of genes	Metabolic network analysis
Error prone PCR	Metabolic control analysis
Gene shuffling	

Another important tool required for metabolic engineering is access to promoters with varying strength. Often it is of interest to increase expression of a certain gene, and availability of strong promoters is therefore desirable. Glycolytic promoters are generally a good choice for strong, constitutive promoters, e.g., the promoter of the gene encoding glyceraldehyde-3-P dehydrogenase that has been cloned in many different microorganisms. For

many organisms there are also strong regulated promoters available, e.g., the *GAL7* promoter of *S. cerevisiae* and the TAKA-amylase promoter of *Aspergillus oryzae*, but the application of these promoters generally requires more advanced fermentation strategies. For the production of heterologous proteins in *E. coli* the *lac* promoter has often been used, since with this promoter it is possible to induce the expression by addition of isopropyl-β-D-thiogalactose (IPTG).

Very often metabolic engineering is carried out with laboratory strains since they are much easier to work with, especially when it comes to genetic engineering. Here it is possible to introduce specific genetic changes, and this enable comparison of strains that are otherwise isogenic. For this purpose very specific disruption cassettes have been developed in certain organisms, e.g., the *loxP-kanMX-loxP* disruption cassette for gene disruption in *S. cerevisiae* (see Fig. 3A). The advantage of this cassette is that the dominant marker can be looped out, and thereby the cassette can be used for sequential disruption of several genes. In *S. cerevisiae* there is a very high frequency of homologous recombination, and only short fragments (down to 50 base pairs) of the genes are required for flanking the resistance marker. For many other microorganisms the frequency of homologous recombination is much lower, and larger fragments are required [for filamentous fungi fragments up to 5 kilobase (kb) may be needed]. Some industrial strains are polyploid, and it is therefore laborious to disrupt genes on all the chromosome pairs. In principle the *loxP-kanMX-loxP* disruption cassette can be used to disrupt the same gene on several chromosomes by repetitive use. Alternatively, disruption of the same gene present on several chromosomes can be done by isolation of haploid spores, disruption of the genes in these haploid strains (or isolation of haploid strains with the proper gene disruption), and subsequent reisolation of a polyploid strain through sexual crossing of the haploid strains. With this procedure it is, however, almost impossible to introduce specific genetic changes without altering the overall genetic makeup of the cell, and the effect of a gene disruption can be difficult to evaluate. An alternative strategy for silencing of gene expression is to apply RNA-antisense techniques (Fig. 3B). If the antisense gene fragment is expressed from a strong promoter, there may be produced sufficient antisense mRNA to silence expression from several gene copies, and this technique is therefore attractive in polyploid industrial strains. Normally, strategies for metabolic engineering developed in laboratory strains can easily be transferred to industrial strains, but in some cases a successful strategy in a laboratory strain will not work in an industrial strain due to a very different genetic background. It is therefore often necessary to evaluate strategies in both laboratory and industrial strains.

With the rapid development in techniques for gene cloning, it has become possible to recruit genes from many different organisms. The increasing access to new

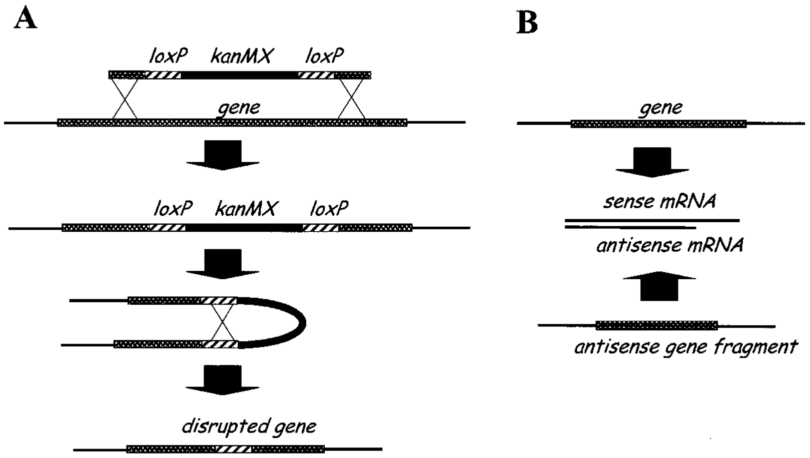


FIGURE 3 Methods for silencing gene expression. (A) Gene disruption by loop-in-loop-out method. Within a part of the gene is cloned a resistance marker (e.g., the *kanMX* gene that ensures resistance towards the antibiotic G418 in *S. cerevisiae*). The resistance marker is flanked by two directed repeats (or *loxP* sequences) and fragments of the gene to be disrupted. Upon transformation there may occur homologous recombination, and the resistance marker is inserted within the gene, which hereby becomes disrupted. The insert can be looped out again through crossover between the two directed repeats. The end result is a disrupted gene, and since the resistance marker has been looped out it can be used again for another disruption. (B) Silencing of gene expression by expression of an antisense mRNA fragment. A part of the gene is inserted in the opposite direction (often behind a strong promoter), and upon expression an antisense mRNA is formed. This may hybridize with the sense mRNA and hereby the translation process is prevented.

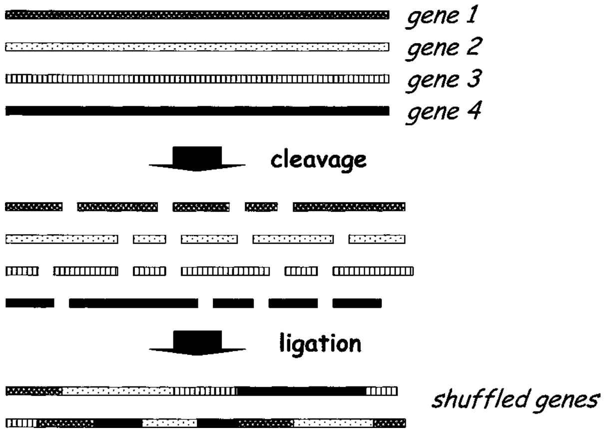


FIGURE 4 Gene shuffling. A family of genes are cleaved by restriction enzymes, and ligated randomly. The shuffled genes are cloned into a proper host, and there are selected for desirable properties. A good selection procedure is clearly essential since a very high number of shuffled genes may be obtained.

organisms through exploitation of the biodiversity in the world will therefore play an important role in the future of metabolic engineering, especially since the access to a class of genes from different organisms will enable construction of completely new genes through gene shuffling (see Fig. 4). Thereby it is possible to construct genes that encode proteins with altered properties, and this approach may also be applied for obtaining enzymes with improved properties, e.g., improved stability, improved catalytic activity, or improved affinity toward the substrate(s). Another approach is to apply directed evolution of a given gene through error-prone polymerase chain reaction (PCR), a technique that may also be used in combination with gene shuffling. Generally these empirical approaches have shown to be more powerful than protein engineering. In the future, when our understanding between the primary structure of proteins and their function has been improved, directed modifications of proteins through protein engineering may, however, enable construction of tailor-made enzymes that may contain desirable properties. Thus, enzymes that have increased affinity for a branch point metabolite may be constructed, and hereby more carbon may be directed toward the end product of interest (see discussion later).

IV. METABOLIC NETWORK ANALYSIS

A key aspect in the field of metabolic engineering is analysis at the cellular level in order to understand the cellular function in detail. Of particular importance is quantification of fluxes through the different metabolic pathways and analysis of how these fluxes are controlled. The metabolic fluxes represent a very detailed phenotypic characterization, and the *in vivo* fluxes are the end result of many different types of regulation within the cell (see Fig. 5). In recent years some very powerful techniques have been developed for quantification of metabolic fluxes and for identification of the active metabolic network—often referred to as metabolic network analysis. Metabolic network analysis basically consists of two steps:

- Identification of the metabolic network structure (or pathway topology)
- Quantification of the fluxes through the branches of the metabolic network

For identification of the metabolic network structure, one may gain much information through the extensive biochemistry literature and biochemical databases available on the web (see, e.g., www.genome.ad.jp, which gives complete metabolic maps with direct links to sequenced genes and other information about the individual enzymes). Thus, there are many reports on the presence of specific enzyme activities in many different species, and for most industrially important microorganisms the major metabolic routes have been identified. However, in many cases the complete metabolic network structure is not known, i.e., some of the pathways carrying significant fluxes have not been identified in the microorganism investigated. In these cases enzyme assays can be used to confirm the presence of specific enzymes and determine the cofactor requirements in these pathways, e.g., whether the enzyme uses NADH or NADPH as cofactor. Even though enzyme assays are valuable for confirming the presence of active pathways, they are of limited use for identification of pathways in the studied microorganism. For these purposes, isotope-labeled substrates is a powerful tool, and

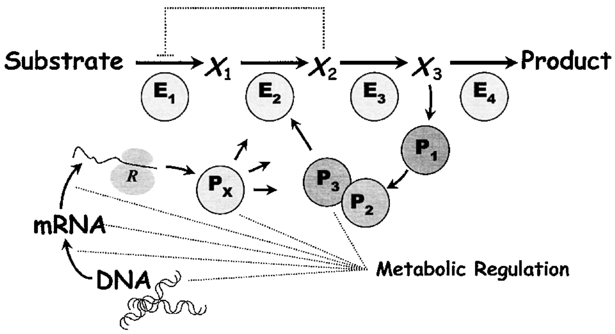


FIGURE 5 Control of flux at different levels. The transcription of genes to mRNA is controlled, and together with control of mRNA degradation this determines the mRNA levels in the cell. The mRNAs are translated into proteins, either enzymes catalyzing biochemical reactions or regulatory proteins acting as transcriptional factors or protein kinases. Finally, the enzymes catalyzing biochemical reactions determine the levels of the metabolites, which influence the metabolic fluxes directly or indirectly through feedback interaction with regulatory proteins. Thus the metabolites indirectly may control both transcription and translation.

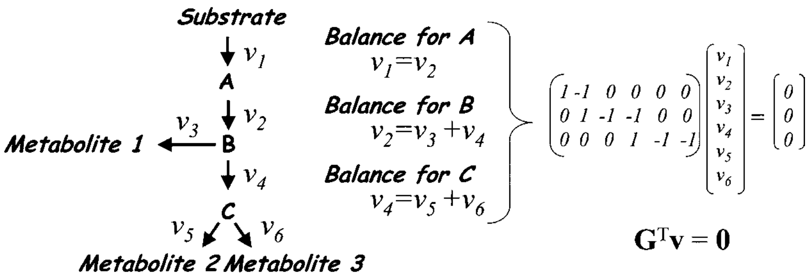


FIGURE 6 Quantification of metabolic fluxes by metabolite balancing. It can generally be assumed that inside the cell the formation and consumption of metabolites is balanced (only immediately after large perturbations does the metabolite concentration change). This gives a set of constraints on the fluxes, and this can be generalized to the matrix equation specified in the figure. In the example there are three equations, and with six fluxes it is possible to calculate three fluxes if three are measured, i.e., the degrees of freedom is three. Note that the balance for metabolite A is obvious, and for this reason linear segments in the metabolic network are normally lumped into overall reactions. In many cases cofactors impose additional constraints between the fluxes, and thus the degrees of freedom may be further reduced.

especially the use of ^{13}C-labeled glucose and subsequent analysis of the labeling pattern of the intracellular metabolites has proven to be very useful for identification of the metabolic network structure. The labeling pattern of ^{13}C in intracellular metabolites may be analyzed either using NMR or using gas chromatography–mass spectroscopy (GC-MS), with the latter technique being superior due to its high speed and sensitivity.

When the metabolic network structure has been identified, it is important to quantify the fluxes through the different branches in the network. The simplest approach to quantify the fluxes is by using the concept of metabolite balancing (see Fig. 6). Here material balances are set up over each metabolite in the network structure, and assuming steady state in the metabolite concentrations a set of algebraic equations relating the fluxes is obtained. These equations impose a set of constraints on the fluxes through the individual reactions in the network. By measuring some of the fluxes or by using linear programming, it is then possible to calculate the fluxes through all the branches of the network. Notice that cofactors may link the individual pathway segments, and thus impose additional constraints on the fluxes. Due to its simplicity, the concept of metabolite balancing is attractive, but it has some limitations. Thus, the flux estimates depend on the cofactor balances, i.e., the balances for NADH and NADPH, and it is therefore important that all reactions involving these cofactors within the cell are included. Since it is unlikely that all reactions involving these cofactors have been identified, metabolite balancing may result in poor estimates of some metabolic fluxes.

Through the use of ^{13}C-labeled glucose and measurement of the labeling pattern of the intracellular metabolites by NMR or GC-MS, it becomes possible to apply balances for the individual carbon atoms in addition to the metabolite balances. Therefore an additional set of constraints is obtained, and it is therefore not necessary to include balances for cofactors. Furthermore, since many balances for the individual carbon atoms can be applied, an overdetermined system of equations is obtained, i.e., there are more equations than unknown fluxes. This redundancy in the equation system enables a more robust estimation of the fluxes, and it also enables estimation of reversible fluxes in the network. Clearly the use of labeled substrates enables a much better estimation of the metabolic fluxes, but it is also a more complex procedure. First of all, measurement of the labeling pattern of the intracellular metabolites requires more advanced analytical procedures, but the equation system is also far more complicated. In recent years this approach has been demonstrated, however, to work very well for estimation of the fluxes in many different microorganisms.

Metabolic network analysis is clearly a very powerful tool for phenotypic characterization. It is, however, important to underline that the technique has no predictive power. Only in few cases do the estimated fluxes by itself point to a strategy for directed genetic modifications. In most cases flux analysis is only useful when different strains are compared or there is performed a comparison of the same strain grown at different environmental conditions (see Fig. 7). Through such comparisons it may be possible to derive correlations between the productivity and certain fluxes, and from such correlations a hypothesis about possible limitations within the cell may be derived.

V. METABOLIC CONTROL ANALYSIS

When the fluxes through the different branches of the metabolic network have been quantified, the next question

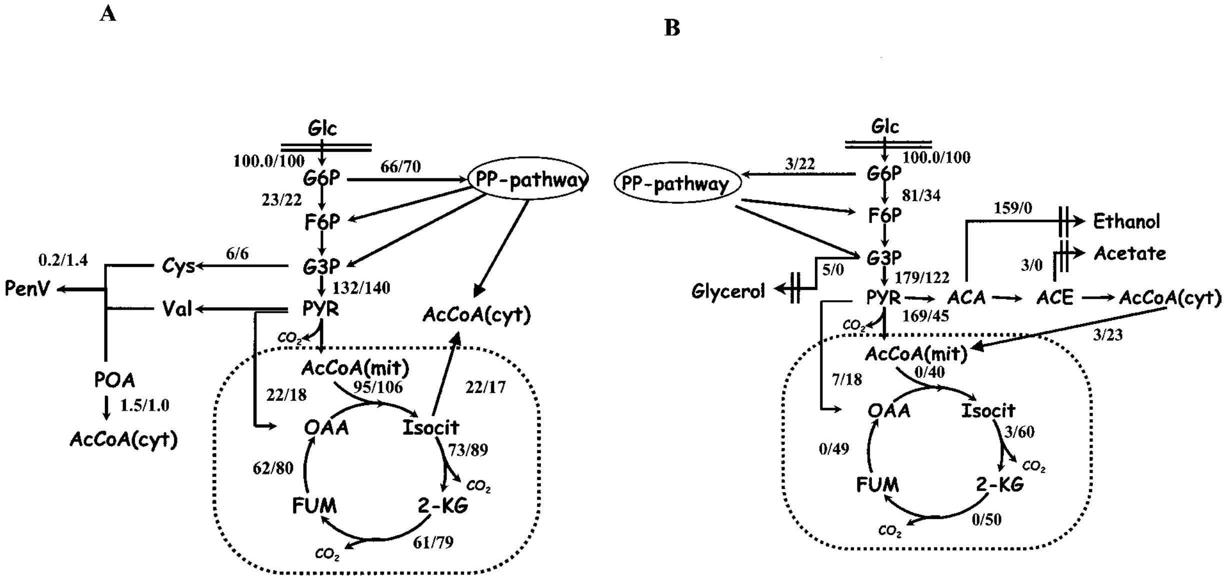


FIGURE 7 Metabolic fluxes estimated using ^{13}C-labeled glucose. (A) Fluxes estimated in a high and low yielding strain of *P. chrysogenum*. The left figures indicate fluxes in the low yielding strain and the right figures indicate the fluxes in the high yielding strain. It is observed that there is a slightly higher flux through the pentose phosphate pathway in the high yielding strain, which may be explained by the increased demand for NADPH in this strain. This points to a possible correlation between penicillin production and pentose phosphate pathway activity. [The data are taken from Christensen, B., Thykær, J., and Nielsen, J. (2000). *Appl. Microbiol. Biotechnol.* **54,** 212–217.] (B) Fluxes estimated in *S. cerevisiae* grown at respectively high and low specific glucose uptake rates. The fluxes at high specific glucose uptake rates are the left figures and the fluxes at low specific glucose uptake rates are the right figures. At high specific glucose uptake rates there is ethanol formation due to the Crabtree effect, and respiration is repressed resulting in no flux through the TCA cycle. Due to the Crabtree effect, the yield of biomass on glucose is low and there is therefore a low requirement for NADPH and precursors of the pentose phosphate pathway, and the flux through this pathway is therefore low. [The data are taken from Gombert, A. K., dos Santos, M. M., Christensen, B., and Nielsen, J. (2001). *J. Bacteriol.* **183,** 1441–1451.]

arises: How is the distribution of flux controlled? This is the key question in metabolic engineering, since it is only when an understanding of how the flux distribution is controlled that one is able to design a suitable strategy to modulate the flux such that an improved cellular performance is obtained. Control of fluxes is determined by kinetic and thermodynamic constraints, and the material balances used for calculating the fluxes therefore supply no information about this. In order to understand the flux control, it is necessary to understand how the enzymes around the branch points in the metabolic network are regulated. Furthermore, it is important to have information about the metabolite levels, which together with information about the affinities of the enzymes in the pathway supply valuable information about the *in vivo* regulation. This can be illustrated by the simple pathway structure in Fig. 8. The distribution of flux through the two branches is determined by three factors:

- The enzyme concentrations
- The affinities of the enzymes toward the metabolite I
- The concentration of the metabolite I

In order to gain information about flux control, methods for measurement of the intracellular metabolites are therefore valuable. Due to the rapid turnover of intracellular metabolites, there are basically two requirements for reproducible analysis of intracellular metabolites: (1) a method for rapid quenching of the cellular metabolism, and (2) efficient analytical procedures that enable measurement in a complex matrix. Using rapid sampling in,

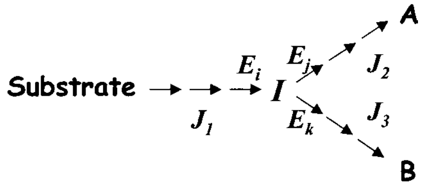


FIGURE 8 Simple pathway structure that illustrates that flux distribution around the branch point metabolite I is controlled by the enzyme concentrations, the affinities of the enzymes for the metabolite, and the metabolite concentration. J_1, J_2, and J_3 represent the steady state fluxes through the three branches of the pathway.

e.g., cold methanol, another cold buffer, or boiling ethanol it is possible to obtain a very rapid inactivation of the cellular metabolism. Using enzymatic assays or different chromatographic techniques, it is possible to measure many different metabolites both in complex matrices and with a high sensitivity, and especially the increased sensitivity of analytical procedures has been of importance for reproducible analysis of intracellular metabolites.

For quantification of flux control, the concept of metabolic control analysis (MCA) is useful. In MCA flux control is quantified in terms of the so-called flux control coefficients (FCCs). The FCCs quantify the relative increase in a given flux J_j within the network upon an increase in a given enzyme activity (E_i), and they are mathematically defined as

$$C_i^{J_j} = \frac{E_i}{J_j}\frac{\partial J_j}{\partial E_i}. \qquad (1)$$

Besides the FCCs there is another set of parameters that are used to characterize the system, namely the elasticity coefficients, which are given by

$$\varepsilon_{X_j}^{i} = \frac{X_j}{v_i}\frac{\partial v_i}{\partial X_j}. \qquad (2)$$

The elasticity coefficients specify the sensitivity of the individual enzymatic reactions to changes in the metabolite concentrations. Thus, if an enzyme is saturated it is clearly not very sensitive to changes in the metabolite concentration, and the elasticity coefficient is low, whereas if the enzyme is not saturated with the metabolite the reaction is sensitive toward changes in the metabolite concentration, i.e., the elasticity is high. The FCCs and the elasticity coefficients are related to each other via the so-called summation theorem, which states that the sum of all the FCCs is 1, and the connectivity theorem, which states that the sum of the product of the elasticity coefficients and the FCCs is zero. If the elasticity coefficients are known, it is therefore possible to calculate the FCCs.

There are different experimental methods available for determination of the FCCs, and these can be grouped into two:

- Direct methods, where the control coefficients are determined directly
- Indirect methods, where the elasticity coefficients are determined and the control coefficients are calculated from the theorems of MCA

Table II gives and overview of the different direct and indirect methods.

Whereas the elasticity coefficients are properties of the individual enzymes, the FCCs are properties of the system. The FCCs are therefore not fixed but change with the environmental conditions, as illustrated in Fig. 9, which summarize results from analysis of the flux control in the penicillin biosynthetic pathway. The penicillin biosynthetic pathway consists of three enzymatic steps. In the

TABLE II Overview of Methods for Determination of FCCs

Method	Procedure	Advantages/disadvantages
Direct		
Genetic manipulations	Alternate the expressed enzyme activity through genetic manipulations, e.g., insert inducible promoters	Robust method that give direct answers, but the method is very laborious
Enzyme titration	Vary the enzyme activity through titration with purified enzymes	Simple and straightforward procedure, but it can only be applied for pathway segments that are completely decoupled from the rest of the cell
Inhibitor titration	Vary the enzyme activity through titration with specific inhibitors	Simple and easy to apply, but requires the existence of specific inhibitors
Indirect		
Double modulation	Measure the metabolite levels at different environmental conditions and determine the elasticity coefficints by calculation of differentials	Elegant approach, but requires two independent changes in the metabolite levels, which is difficult to obtain due to the high degree of coupling between intracellular reactions
Single modulation	Similar to double modulation but based on knowledge of one of the elasticity coefficients	More robust than double modulation, but it requires knowledge of one elasticity coefficient
Top-down approach	Based on grouping of reactions and then using, e.g., double modulation	Very useful, but do not directly give all the FCCs of the system
Kinetic models	Direct calculation of the elasticity coefficients from a kinetic model	Robust, but relies on the availability of a reliable kinetic model for the individual enzymes in the pathway

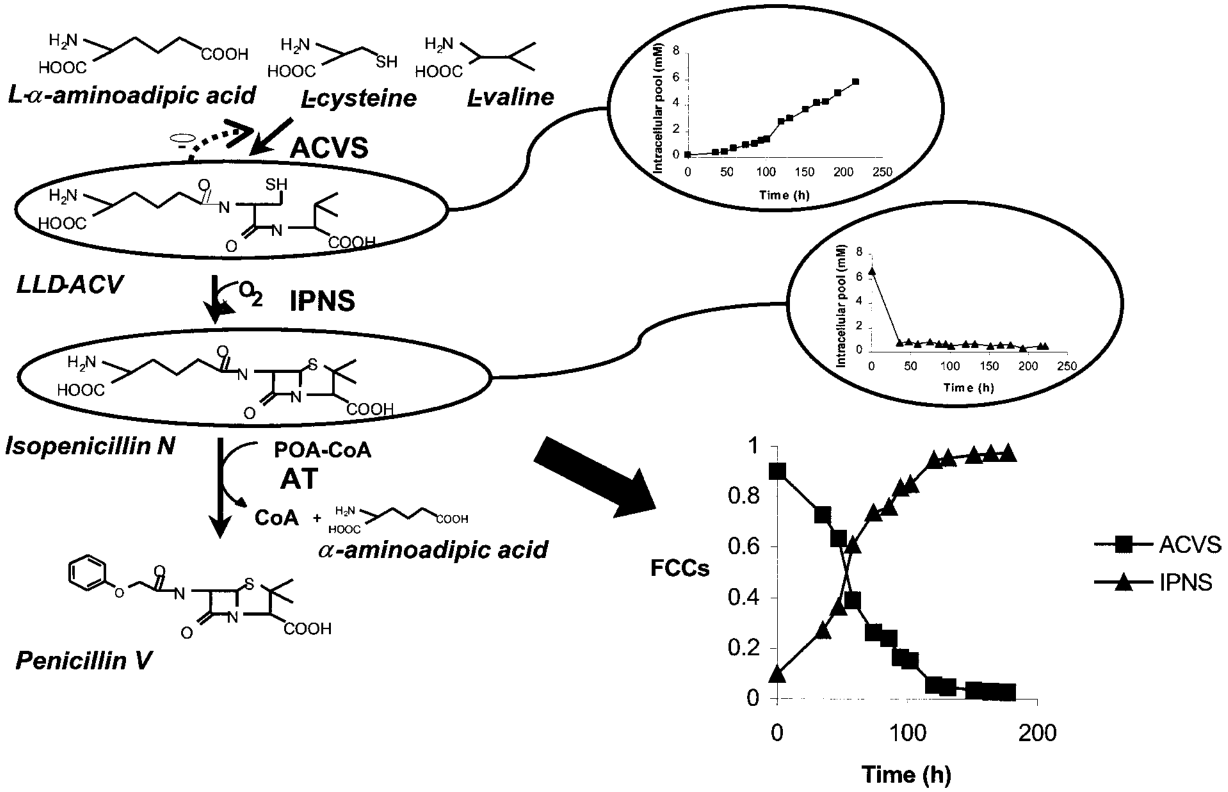


FIGURE 9 MCA of the penicillin biosynthetic pathway. Based on a kinetic model for the enzymes, in this pathway the FCCs were calculated at different stages of fed-batch cultivations. During the first part of the cultivation the flux control was mainly exerted by the first step in the pathway, i.e., the formation of the tripeptide LLD-ACV by ACV synthetase (ACVS), whereas later in the cultivation flux control shifted to the second step in the pathway, i.e., the conversion of LLD-ACV to isopenicillin N by isopenicillin N synthetase (IPNS). This shift in flux control is due to intracellular accumulation of LLD-ACV, which is an inhibitor of ACVS. The initial high isopenicillin N concentration is due to the fact that this sample was taken from the inoculum culture where the side-chain precursor phenoxyacetic acid (POA) was not present in the medium. [The data are taken from Nielsen, J., and Jørgensen, H. S. (1995). *Biotechnol. Prog.* **11,** 299–305.]

first step the three amino acids L-α-aminoadipic acid, L-valine, and L-cysteine are condensed into the tripeptide δ-(L-α-aminoadipyl)-L-cysteinyl-D-valine, normally abbreviated ACV. This reaction is catalyzed by ACV synthetase (ACVS), which besides formation of the two peptide bonds also performs the epimerisation of the valine residue. The ACVS is feedback inhibited by ACV. In the second step, ACV is converted to isopenicillin N—a reaction catalyzed by isopenicillin N synthase (IPNS). This reaction is unique because oxygen is used as electron acceptor. In the last reaction the side chain of isopenicillin N is exchanged with phenoxyacetic acid (POA), resulting in the formation of penicillin V. This reaction is carried out by acyltransferase (AT). Before incorporation into penicillin V, the side-chain precursor phenoxyacetic acid has to be activated as a CoA-ester, which is performed by a specific CoA ligase. Based on a kinetic model, the elasticity coefficients were derived and hereby the FCCs could be determined by an indirect method. The elasticity coefficients are functions of the intracellular metabolite concentrations, and the resulting FCCs are therefore also functions of the concentration of the intermediates of the pathway. From analysis of the pathway intermediates during fed-batch fermentations it was found that ACV accumulated during the fermentation whereas the isopenicillin N was approximately constant during the fermentation (see Fig. 9). Initially the ACVS is the flux-controlling enzyme (with an FCC close to 1), but due to the feedback inhibition of ACV on the first enzyme, flux control shifts during the fermentation to IPNS, which in the later part of the fermentation is the flux-controlling enzyme (with an FCC close to 1). Obviously, it makes no sense to talk about a "rate-limiting step" or a "bottleneck enzyme" in this process since the flux control shifts during the process.

VI. TOOLS FROM FUNCTIONAL GENOMICS

The active metabolic network functioning in a given cell is determined by the enzymatic makeup of the cell. Fluxes are therefore indirectly also controlled at the level of transcription and translation. In fact, there are several different levels of control as illustrated in Fig. 5:

1. Transcriptional control
2. Control of mRNA degradation
3. Translational control
4. Protein activation/inactivation
5. Allosteric control of enzymes

Due to this hierarchical control, it is difficult to predict the overall consequences of a specific genetic modification. Thus, a genetic change may result in altered enzyme levels, and thereby the metabolite concentrations may change. This may influence regulatory proteins that may lead to a secondary effect, both at the transcriptional level and at the level of the enzymes. Generally, it is difficult to predict all the consequences of a certain specific genetic change, and it may therefore be necessary to go through the cycle of metabolic engineering several times. However, with novel analysis techniques developed within the field of functional genomics, a very detailed characterization of the cell can be carried out, and this may enable a far better design of a metabolic engineering strategy.

A very powerful analytical technique that has emerged in recent years is DNA arrays or DNA chips, which enables measurement of the expression profile of all genes within a genome—often referred to as whole genome transcriptome analysis. Thus, with one measurement it is possible to monitor which genes are active at specific conditions, and especially it is possible to rapidly screen single deletion mutants. Combination of DNA arrays for genome wide expression monitoring and bioinformatics has demonstrated to be very powerful for pathway reconstitution, and in the future this approach is expected to speed up the assignment of function to orphan genes. However, it is also expected to play a very valuable role in metabolic engineering.

Since the introduction of the concept of DNA arrays in the early nineties, the technology has developed rapidly. Although DNA arrays are still in their infancy, they have already been applied to many different applications. In April 1996 the American company Affymetrix launched the first commercial DNA array product, a chip designed to look for mutations in the HIV genome. Chips for mutation analysis in the p53, p450, and BRCA1 genes soon followed. Besides mutation analysis chips, chips have become available for gene expression analysis in different organisms, and today there are commercial available chips for *E. coli*, *S. cerevisiae*, plant cells, different animal cells (only part of the genome currently covered), and human cells (also only with part of the genome currently covered). These chips contain oligonucleotide sequences from sequenced genes and allow rapid identification of complementary sequences in the sample mRNA pool (see Fig. 10). DNA arrays produced by Affymetrix are generated by a photolitography process, which allows the synthesis of large numbers of oligonucleotides directly on a solid substrate (see www.affymetrix.com for further details). An alternative method for production of DNA arrays is spotting of oligonucleotide or cDNA solutions by a robot to a solid substrate followed by immobilization of the oligonucleotide or DNA. This method allows for production of custom-made DNA arrays, and is therefore more flexible. Normally it is cDNA that is spotted on these custom-designed arrays, and the cDNA may have a length of several hundred nucleotides. This enables a very good hybridization, and it is therefore not necessary to have more than a single probe for each gene.

Besides DNA arrays for transcription profiling, other tools from functional genomics are valuable in the field of metabolic engineering. Using two-dimensional electrophoresis, it is possible to identify the pool of proteins present in a given cell—often referred to as proteomics—

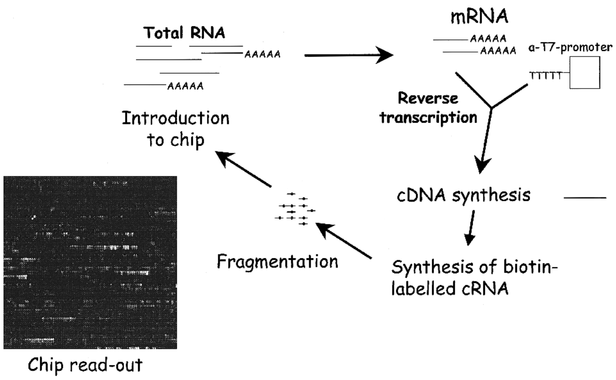


FIGURE 10 Procedure for analysis of transcription profiles of eucaryotes using a DNA chip from Affymetrix. The total RNA is extracted from the cell and the mRNA is purified. All eucaryotic mRNA contains a polyA tail, and this can be used to synthesize cDNA by reverse transcription. The cDNA is synthesized with a T7 promoter, which enables later synthesis of biotin-labeled cRNA. The biotin-labeled cRNA is fragmented and introduced into the chip. The cRNA fragments then hybridize to the oligonucleotides in the chip, and the concentration of the cRNA fragments can be analyzed by a fluorescence scanner as illustrated in the chip readout. For each gene there are 15–20 different oligonucleotides (or probes) of a length of 20–25 nucleotides each.

and combined with transcription profiling this may give important information about regulation at the different levels. Finally, powerful analytical techniques like GC-MS and LC-MS-MS (LC: liquid chromatography) enable measurement of a large fraction of the intracellular metabolite—often referred to as the metabolome. As discussed above, the intracellular metabolite concentrations give unique information about the control of fluxes, and therefore represent a very detailed phenotypic characterization. There is still, however, some significant developments required before these methods enable high-throughput, quantitative analysis of many metabolites.

VII. APPLICATIONS OF METABOLIC ENGINEERING

Today metabolic engineering is applied in the optimization of almost all fermentation processes, and the total market value of fermentation-derived products exceeds 55 billion US$. Table III lists some typical fermentation products categorized according to their synthesis route, i.e., whether they are whole cells, derived from the primary or secondary metabolism, are specifically synthesized (and perhaps secreted) proteins, large polymers, or genetic material. Clearly, the strategy for improving the productivity depends much on whether the product is synthesized by the action of many enzymatic steps or whether the product is derived directly from expression of a single gene. Furthermore, for high value added products like pharmaceutical proteins, time to market is often more important than obtaining a high yield or high productivity, which on the other hand are essential for optimization of processes leading to low-value added products like ethanol and many antibiotics. Despite these differences, the mindset of metabolic engineering is still extremely valuable in optimization of any fermentation processes, as discussed further in the following.

In recent years there have been reported on many examples of metabolic engineering, and these examples can be

TABLE III List of Some Fermentation Products and Some Market Volumes

Category of product	Product	Typical organism	
Whole cells	Baker's yeast	*S. cerevisiae*	
	Lactic acid bacteria	Lactic acid bacteria	
	Single cell protein	Methanogenic bacteria	
Primary metabolites	Ethanol	*S. cerevisiae*, *Z. mobilis*	12 billion US$
	Lactic acid	Lactic acid bacteria, *R. oryzae*	200 million US$
	Citric acid	*A. niger*	1.5 billion US$
	Glutamate	*C. glutamicum*	1 billion US$
	Lysine	*C. glutamicum*	500 million US$
	Phenylalanine	*E. coli*	200 million US$
Secondary metabolites	Penicillins	*P. chrysogenum*	4 billion US$
	Cephalosporins	*A. chrysogenum*, *S. clavuligerus*	11 billion US$
	Statins	*Aspergillus*	9 billion US$
	Taxol	*Plant cells*	1 billion US$
Recombinant proteins	Insulin	*S. cerevisiae*, *E. coli*	3 billion US$
	tPA	CHO cells[a]	1 billion US$
	Erythropoitin	CHO cells	3.6 billion US$
	Human growth hormone	*E. coli*	1 billion US$
	Interferons	*E. coli*	2 billion US$
	Vaccines	Bacteria and yeast	
	Monoclonal antibodies	Hybridoma cells	700 million US$
Enzymes	Detergent enzymes	*Bacillus*, *Aspergillus*	600 million US$
	Starch industry	*Aspergillus*	200 million US$
	Chymosin	*Aspergillus*	
Polymers	Xanthan gum	*X. campestris*	400 million US$
	Polyhydroxybutyrates		
DNA	Vaccines	*E. coli*	
	Gene therapy	*E. coli*	

[a] Chinese hamster ovary cells.

divided into seven categories, depending on the approach taken or of the aim:

- Heterologous protein production
- Extension of substrate range
- Pathways leading to new products
- Pathways for degradation of xenobiotics
- Engineering of cellular physiology for process improvement
- Elimination or reduction of by-product formation
- Improvement of yield or productivity

Below follows a short discussion of the different categories with some presentation of a few examples.

A. Heterologous Protein Production

The first breakthrough in genetic engineering paved the way for a completely new route for production of pharmaceutical proteins like human growth hormone (hGH) and human insulin; it also opened the possibility to produce many other pharmaceuticals. The first products (human insulin and hGH) were produced in recombinant *E. coli*, but soon followed the exploitation of other expression systems like *S. cerevisiae* (introduced for production of human insulin), insect cells, and mammalian cells (Chinese hamster ovary cells and hybridoma cells). Today there are more than 55 protein drugs, largely recombinant proteins and monoclonal antibodies that are often referred to as biotech drugs, and the 20 top-selling drugs represents sales of more than 16 bio US$. The choice of expression system depends upon many factors:

1. the desirability of posttranslational modification and secretion
2. the stability of the protein in question
3. the projected dose of protein per patient (which determines whether the cost of the drug becomes critical)

Thus for proteins used in large doses, like human insulin, it is important that the production costs are kept low, which requires an expression system with a high productivity, i.e., *E. coli* or *S. cerevisiae*. For very complex molecules like tissue plasminogen activator (tPA) and erythropoietin, it is not, however, possible to obtain sufficiently active compounds in microbial systems, and here a higher eukaryotic expression system is required.

When a certain expression system has been implemented for the production of one product, it is often desirable to use this expression system within the company to produce other products as well. There have therefore been several attempts to engineer generic expression systems for improved performance for heterologous protein production. A major problem encountered with this strategy is, however, that the secretory capacity in a given expression system is strongly protein specific. Thus, for one protein it may be a certain step in the secretion pathway that is limiting the production, e.g., folding, formation of disulfide bridges, or glycosylation, whereas other proteins may easily be processed through the pathway and secreted at high rates. Even when working with generic expression systems it is therefore normally necessary to further optimize the expression system for the specific protein to be produced, and there have been several attempts to engineer the glycosylation pathway or the secretion pathway.

B. Extension of Substrate Range

Biotech processes are attractive for replacement of many classical chemical processes, since they potentially may apply agricultural waste products as raw materials (typical as carbon source), and thereby sustainable industrial processes may be obtained for the production of fuels, chemicals, and materials. Often the industrial strain applied for a given process has a narrow substrate spectrum, and it is therefore necessary to extend the substrate range. Here it is relevant to consider two different strategies:

- Introduction of a gene encoding a membrane-bound protein that transports the substrate into the cell in addition to genes encoding the necessary pathway that is responsible for channeling the substrate into the central carbon metabolism
- Introduction of a gene encoding a secreted protein that converts the substrate to compounds that can be directly assimilated and metabolized by the host organism

Expression of specific permeases is difficult, but the first strategy is still the preferred one for engineering cells to take up monosaccharides since often these compounds can be taken up by nonspecific permeases. Thus, one can focus on engineering the pathway converting the monosaccharide or disaccharide into the central carbon metabolism. The second strategy is typically applied for di-, oligo-, and polysaccharides, which may be difficult (or impossible) to transport into the cell. Here it is necessary to ensure efficient secretion of an enzyme (often of heterologous nature) that may degrade the substrate. Besides ensuring proper secretion of the hydrolytic enzyme, it is important to ensure that the hydrolysis rate is sufficiently high to ensure supply of mono- or disaccharides for growth. In the literature there are many examples of extension of the substrate range and among the most prominent example is engineering of *S. cerevisiae* such that it can metabolize

xylose, which is present in plant material as part of the heterogeneous polymer hemicellulose. Hemicellulose is one of the main constituents of lignocellulose, which due to its abundance and low cost is an attractive raw material for low value added products, e.g., fuel ethanol.

C. Pathways Leading to New Products

Metabolic engineering offers immense possibilities for constructing pathways leading to novel products, and there are many beautiful examples in the literature. There may be three different objectives for extending pathways to produce new products in a certain organism:

- Production of completely new compounds: This is of particular importance in the field of pharmaceuticals, where artificial pathways may produce new antibiotics and anticancer drugs. However, also in the field of materials it may be possible to tailor-make plastics with specific properties.
- Through introduction of new pathways in a given microorganism, it may become possible to develop biotech-based process that can replace classical chemical processes. This is primarily of interest since biotech processes are more environmentally friendly, but also the selectivity of biochemical reactions may offer the possibility to produce optically pure compounds more efficiently. Since biotech processes use sustainable raw materials and generally are more environmentally friendly than chemical synthesis, these are often referred to as green chemistry.
- Exploitation of a common host for production of many different products: Thus it is possible to capitalize investments in optimizing the fermentation process with the general host. This is especially exploited in the field of enzyme production, where a few host systems are applied to produce a wide range of enzymes. Also, in the field of amino acid production there is focus on applying a few production systems, i.e., *E. coli* and *C. glutamicum*.

One area that illustrates the power of metabolic engineering is the development of alternative routes for the production of 7-amino cephalosporanic acid (7-ACA) and 7-amino deacetoxycephalosporanic acid (7-ADCA), which serves as precursors for the production of semisynthetic cephalosporins. Until recently the sole route for 7-ADCA was by chemical ring expansion of 6-amino penicillanic acid (6-APA), which can be derived from penicillins. Thus, many semisynthetic cephalosporins were traditionally derived from penicillin produced by fermentation. With the high potential of β-lactam production by *P. chrysogenum*, it is of interest to engineer this organism to produce the cephalosporins directly by fermentation. By transforming a strain of *P. chrysogenum* with genes encoding an expandase from *Streptomyces clavuligerus*, it has been possible to produce adipoyl-7-ADCA directly by fermentation with transformed strains of *P. chrysogenum* (see Fig. 11). Similarly, it is possible to produce adipoyl-7-ACA directly be fermentation with a recombinant strain of *P. chrysogenum* harbouring genes encoding the expandase/hydroxylase and the acyltransferase from *Acremonium chrysogenum* (see Fig. 11). From adipoyl-7-ADCA and adipoyl-7-ACA the compounds 7-ADCA and 7-ACA can easily be synthesised by enzymatic removal of the adipoyl side chain. This process of direct production of 7-ADCA and 7-ACA by fermentation clearly demonstrates the potential of metabolic engineering in the design of new processes, which is economically more efficient and more environmentally friendly than the chemical synthesis route.

D. Pathways for Degradation of Xenobiotics

Bioremediation in the field of environmental cleanup has attained much attention since the late 1980s. Bioremediation refers to the use of natural microorganisms for remediation of polluted air, water, soil, or sediment. Many field tests and laboratory experiments have identified microorganisms that can degrade harmful organic compounds—often referred to as xenobiotics—like aromatics (e.g., benzene, toluene, xylene), halogenated aromatics (e.g., polychlorinated biphenyls), halogenated aliphatics, and pesticides. However, there are two major problems in connection with wide range application of these microorganisms for biodegradation of xenobiotics:

- The rate of degradation is slow. This is mainly due to the fact that contaminants are often distributed over a wide range, and are present in too low concentrations to induce the degradation pathways.
- Degradation of mixtures of xenobiotics requires several different microbial species.

Metabolic engineering may be used to solve these problems, and it furthermore offers the possibility to construct completely novel xenobiotics degrading pathways through recruitment of enzymes from different organisms.

E. Engineering of Cellular Physiology for Process Improvement

In the industrial exploitation of living cells, their properties may be undesirable, and it is therefore of interest to obtain cells with improved properties through metabolic engineering of the overall physiology of the cell. Some typical problems encountered are as follows:

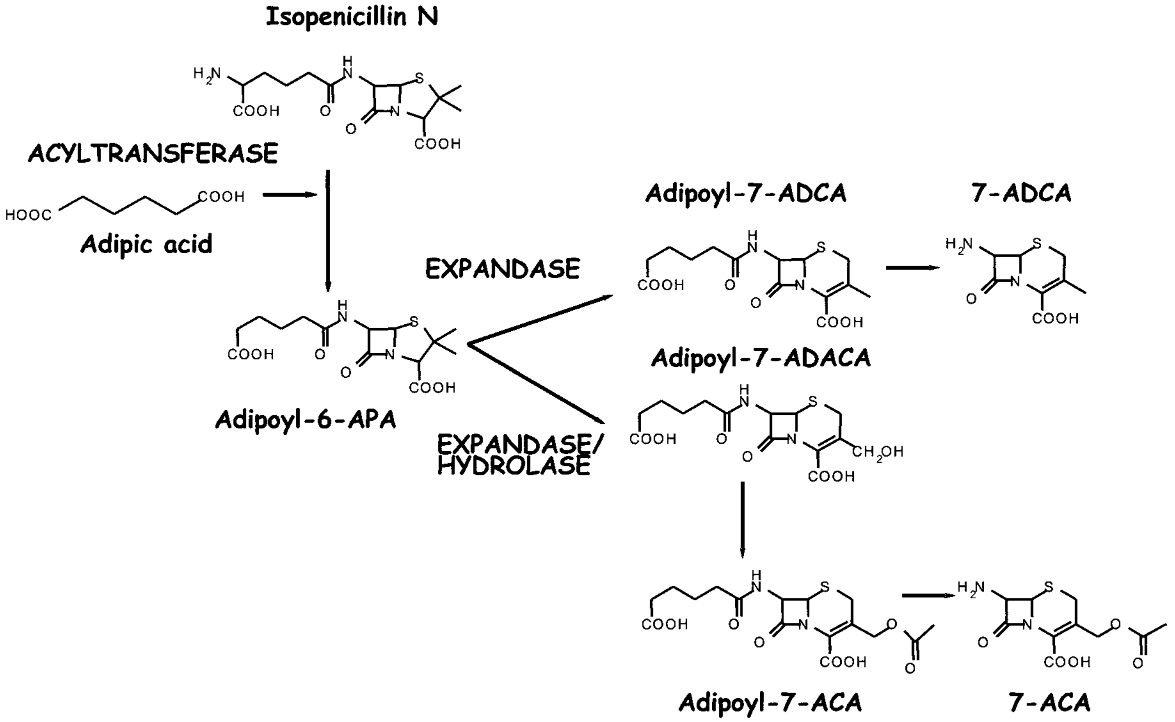


FIGURE 11 Pathways for the production of 7-ADCA and 7-ACA directly by fermentation. In this process the acyltransferase converts isopenicillin N to adipoyl-6-APA, and when the expandase of *S. clavuligerus* is expressed in *P. chrysogenum*, this compound is directly converted to adipoyl-7-ADCA. When *P. chrysogenum* is transformed with the expandase/hydroxylase and the acyltransferase of *Acremonium chrysogenum*, adipoyl-6-APA is converted first to adipoyl-7-ADACA, which is further converted to adipoyl-7-ACA.

- *Glucose repression.* Many industrial fermentations are carried out on sugar mixtures, e.g., molasses that contain sucrose, glucose, fructose, and raffinose, and the presence of glucose represses the utilization of the other sugars. This causes lag phases leading to prolonged fermentation times. Another problem may be encountered in the field of enzyme production, where glucose repression of the expression promoter results in reduced expression of the gene encoding the product. Disrupting DNA-binding proteins that mediates glucose repression in different microorganisms has solved problems with glucose repression.
- *High sensitivity to low oxygen concentrations.* Many microbial cells are sensitive to low oxygen concentration, which may result in reduced product formation—or even irreversible loss of product formation—or onset of fermentative metabolism. This problem can partly be overcome by expression bacterial hemoglobin, which has demonstrated to have a positive influence in many different fermentation processes.
- *Modulation of macroscopic structures.* In some fermentations it is desirable to improve flocculation, e.g., in beer fermentation, whereas in other fermentations it is optimal to have disperse cellular structures, e.g., in fermentations with filamentous microorganisms.

F. Elimination of By-Product Formation

Many fermentation processes have been optimized such that the desired product is predominantly formed. However, due to the complexity of cellular metabolism it is inevitable that by-products are formed. This may be undesirable for at least three reasons:

- The by-product(s) may be toxic to humans or animals.
- The by-product(s) may cause problems in the subsequent separation process.
- Formation of the by-product(s) results in a loss of carbon, and the overall yield of product on the raw material is therefore below the theoretical maximum.

The first reason is clearly problematic if humans or animals may be exposed to the product, either directly or indirectly. Typically, one chooses a cellular system that do not produce toxins when products are made for human

consumption, e.g., fungal cells that do not produce aflatoxins are used for production of food grade enzymes, and *E. coli* strains that do not produce endotoxins are used for production of pharmaceuticals. The second reason may especially be problematic if the by-product has properties similar to those of the desired product, since separation of the active product then requires very efficient separation principles, which are often costly. In some cases the by-product may also cause other types of problems, e.g., inactivation of the desired product. Loss of carbon in the by-product is mainly a problem in connection with the production of low value added products, and here the strategies discussed in the following section may be applied.

G. Improvement of Yield or Productivity

In the production of low value added products like ethanol, lactic acid, citric acid, amino acids, and many antibiotics, the yield of product on the substrate and the productivity are the most important design variables to optimize. Yield impacts primarily the cost of raw materials and is affected by the metabolic fluxes, i.e., an increase in the yield is obtained through redirection of the fluxes toward the desired product. Productivity, on the other hand, is of importance when the capital costs are of importance, and it can be improved by amplification of the metabolic fluxes. In order to obtain a high yield and productivity it is necessary to direct the carbon fluxes from the substrate toward the metabolite of interest at a high rate. This often requires engineering of the central carbon metabolism, which is difficult due to the tight regulation in this part of the cellular metabolism.

For most processes it is not possible to operate the process such that both the yield of product on the substrate and the productivity are at their optimum (see Fig. 12). It is therefore necessary to evaluate which of the two variables is most important. Similarly, an improvement in the productivity through metabolic engineering does not necessarily lead to an improvement in the yield of product on the substrate. The simple model pathways shown in Fig. 8 illustrate this. Here the substrate can be converted both to the product of interest and to a by-product—in practice there is often several different by-products. In this pathway the yield is given as the ratio between the fluxes J_2 and J_1. Clearly the yield is not changed if the flux distribution around the branch point metabolite I is constant, i.e., independently of the size of flux J_1 the ratios J_2/J_1 and J_3/J_1 are constant, whereas the productivity clearly increases when J_1 increases. The flux distribution around the branch point metabolite I depend on the flexibility or rigidity of the network. The flexibility of a branch point is determined by the affinities of the enzymes competing for this metabolite and the concentration of the metabolite. If the metabolite concentration is much below the K_m values of the two enzymes, then the flux distribution is likely to be maintained if the supply of the metabolite via flux J_1 increases. However, if both enzymes are saturated with the metabolite, then an increase in the flux J_1 will have no effect on either the yield or the productivity. In this case one may increase the yield and the productivity through increasing the flux J_2, e.g., by overexpressing the enzyme E_2. However, if an increase in E_j results in a decrease in the concentration of I below the K_m values of the enzymes E_j and E_k, the consequence may be an altered flux distribution. Thus, even in this simple pathway structure it is difficult to design the appropriate strategy for increasing the yield or productivity without knowledge of the kinetics of the enzymes and the metabolite concentration. The complexity increases further if there is feedback inhibition. Here the strategy for improvement

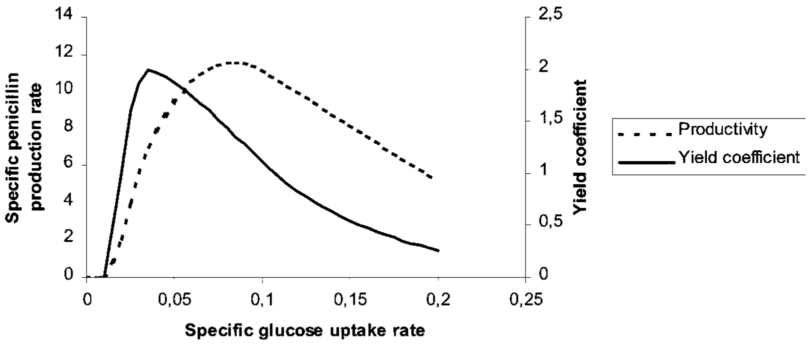


FIGURE 12 The specific penicillin productivity and the yield of penicillin on glucose as function of the specific glucose uptake rate (arbitrary units). The data are typically representatives of penicillin production by *P. chrysogenum* [see, e.g., Henriksen, C. M., Christensen, L. H., Nielsen, J., and Villadsen, J. (1996). *J. Biotechnol.* **45,** 149–164, or van Gullik, W. M., de Laat, W. T. A. M., Vinke, J. L., and Heijnen, J. J. (2000). *Biotechnol. Bioeng.* **68,** 602–618.] Notice that the specific productivity and the yield coefficient have maximum at different specific glucose uptake rates, and it is therefore not possible to optimize both these parameters at the same time.

of yield or productivity depends on the type of inhibition. In the case of competitive inhibition, an increased flux toward the product may result in an increase in the product concentration and thereby a decreased affinity of E_j for the branch-point metabolite. This may lead to a change in flux distribution around the metabolite I—perhaps even a decrease in the overall yield. With noncompetitive inhibition the situation is quite different, since the affinity is not affected by the product concentration. However, in this case an overexpression of E_j may not lead to a significant increase in the *in vivo* activity due to the feedback inhibition. In case of feedback inhibition it is generally the best strategy to increase the step after the inhibitory metabolite as illustrated with the flux control in the penicillin biosynthetic pathway (see Fig. 9). Alternatively, one can introduce feedback-insensitive enzymes, e.g., by gene-shuffling or by recruiting a heterologous enzyme. Furthermore, through gene-shuffling enzymes with reduced or increased affinity may be constructed, and this may enable redirection of pathway fluxes, and hereby a common host can be used for the production of several different metabolites.

VIII. FUTURE DIRECTIONS

Metabolic engineering has come far in the last decade. The first success stories involved introduction of a single heterologous gene for protein production or disruption of a single gene for redirection of pathway fluxes. In recent years, there have also been demonstrations of how multiple genetic modifications can be exploited to obtain the overall goal. These may either come around through several rounds of the metabolic engineering cycle (Fig. 1), or they may be the result of a careful analysis that immediately points to the benefit of introducing several genetic modifications at the same time. The approach of suitably coordinated expression (and/or inhibition) of several genes will be necessary to achieve more advanced objectives, and this calls for a systems approach. Thus, it is necessary to consider the complete metabolic network, or the complete set of signal transduction pathways that are involved in regulation of cellular function, and it is exactly this systems approach that distinguishes metabolic engineering from applied molecular biology. The developments in functional genomics, with new analytical techniques for measurement of the transcriptome using DNA chips, the proteome using 2D gels, and the metabolome using different analytical techniques, will have a significant impact on metabolic engineering. With the tradition of considering whole cell function in the field of metabolic engineering, there are, however, also several lessons from metabolic engineering for functional genomics, and in the future it is expected that these two research areas will benefit much from each other.

ACKNOWLEDGMENT

I dedicate this paper to Jay Bailey, who has served as a constant inspiration for the work carried out in my research group. Jay recently passed away and this will be a great loss to the metabolic engineering community. However, I am sure that his thoughts and ideas will continue to influence this research field in the future.

SEE ALSO THE FOLLOWING ARTICLES

Biomaterials, Synthesis, Fabrication, and Applications • Chromatin Structure and Modification • DNA Testing in Forensic Science • Enzyme Mechanisms • Gene Expression, Regulation of • Hybridomas, Genetic Engineering of • Protein Synthesis • Tissue Engineering • Translation of RNA to Protein

BIBLIOGRAPHY

Bailey, J. E. (1991). Toward a science of metabolic engineering. *Science* **252,** 1668–1674.

Christensen, B., and Nielsen, J. (1999). Metabolic network analysis—Powerful tool in metabolic engineering. *Adv. Biochem. Eng./Biotechnol.* **66,** 209–231.

Fell, D. (1997). "Understanding the Control of Metabolism," Portland Press, London.

Lee, S. Y., and Papoutsakis, E. T. (1999). "Metabolic Engineering," Marcel Dekker, New York.

Ostergaard, S., Olsson, L., and Nielsen, J. (2000). Metabolic engineering of *Saccharomyces cerevisiae. Microbiol. Mol. Biol. Rev*. **64,** 34–50.

Stephanopoulos, G., Aristodou, A., and Nielsen, J. (1998). "Metabolic Engineering," Academic Press, San Diego, CA.

Microanalytical Assays

Jerome S. Schultz
University of Pittsburgh

I. Microfabrication
II. Biosensors
III. Biological Recognition Elements
IV. Immobilization
V. Detector Elements
VI. Optically Based Biosensors
VII. Other Detectors
VIII. Sensor Dynamics

GLOSSARY

Absorbance The common logarithm of the reciprocal of the internal transmittance of a given substance.

Amperometric Titration in which the end point is determined by measuring the amperage of an electric current of given voltage that is passed through the solution.

Analyte A component to be measured in an analysis.

Antibody Protein naturally existing in blood serum or produced by an animal in response to stimulation by an antigen that reacts to overcome the toxic effects of a specific antigen.

Binding The act or process by which one molecule attaches to another by noncovalent forces.

Bioassays Any quantitative estimation of biologically active substances by measurement of the magnitude of their actions on living organisms or parts of organisms.

Biomolecule Any molecule, especially a macromolecule, that occurs in or is formed by living organisms.

Biorecognition An interaction between a bioactive molecule and receptor from a cell.

Biosensor A devise that uses specific biochemical reactions mediated by isolated enzymes, immunosystems, tissues, organelles, or whole cells to detect chemical compounds, usually by electrical, thermal, or optical signals.

Biotechnology The study of the relationship between human beings and machines, especially in terms of physiological, psychological, and technological requirements.

Catalysts Any substance that increases the rate of a chemical reaction but is itself unchanged at the end of the reaction.

Chemiluminescence The production of visible light (luminescence) occurring as a result of a chemical reaction.

Chromophores A molecular moiety that produces color absorption when bound to a colorless molecule.

Electroactive A substance that reacts at an electrode.

Electroosmotic The motion of a liquid through a membrane under the influence of an applied electric field.

Electrophoresis The motion of colloidal particles suspended in a fluid medium, due to the influence of an electric field on the medium. Also called cataphoresis.

Enzyme Any of various complex organic substances, as pepsin, originating from living cells and capable of producing certain chemical changes in organic substances by catalytic action, as in digestion.

Dialysis A process in which solute molecules are exchanged between two liquids through a membrane in response to differences in chemical potentials between the liquids.

Fluorescence A type of luminescence that consists of the emission by a substance of electromagnetic radiation, especially visible light, immediately (10–100 ns) after the absorption of energy derived from exciting radiation of another, usually shorter, wavelength or from incident subatomic particles (especially electrons or α particles); the property of emitting such radiation.

Hybridization The act or process of forming a macromolecular hybrid by the artificial recombination of subunits.

Immobilize To render any agent, whether a micro- or macrosolute, a particle, or an intact cell, nondispersible in an aqueous medium with retention of its specific ligating, antigenic, catalytic, or other properties.

Immunoassays Any of a group of techniques for the measurement of specific biochemcial substances, commonly at low concentrations and in complex mixtures such as biological fluids, that depend upon the specificity and high affinity shown by suitably prepared and selected antibodies.

Immunoglobulin Any member of a group of proteins occurring in higher animals as major components of the immune system.

Macromolecule Any molecule composed of a very large number of atoms, operationally defined as any molecule of mass greater than about 10 kDa that is incapable of passing through the pores of dialysis tubing as generally used.

Osmosis The spontaneous net flow of solvent by diffusion through a semipermeable membrane from a phase where the solvent has a higher chemcial potential to one where the solvent has a lower chemical potential.

Piezoelectric The generation of electricity or of electric polarity in dielectric crystals subjected to mechanical stress and, conversely, voltage.

Polarography An electrochemical method of quantitative analysis based on the relationship between an increasing current passing through the solution being analyzed and the increasing voltage used to produce the current.

Potentiometry A precision instrument for measuring an electric potential difference of constant polarity without drawing current from the circuit being examined.

Receptor Any cellular macromolecule that undergoes combination with a hormone, neurotransmitter, drug, or intracellular messenger to initiate a change in cell function.

Spectrophotometer An apparatus to measure the proportion of the incident light (or other electromagnetic radiation) absorbed (or transmitted) by a substance or solution at each wavelength of the specturm.

Transducer Any device that converts a quantity of energy from one form into another.

ONE OF THE MAJOR technological advances that has helped the field of microassays has been the translation of fabrication methods from the chip industry into microfabrication of devices for chemical sensors. This type of fabrication procedure allows the synthesis of systems that can have various components. Each of these components has some similarities to the components or stages that are used in the traditional chemical analytical laboratory. These stages included sample preparation, sample purification followed by sample measurement, and then finally output of an electronic signal. The microfabrication methods allow the production of multiple array systems that can be as small as 500 nm on each side so that hundreds of the sensors can be placed in one device. A number of transducer technologies can be used as sensors for particular analytes. These include electrodes, optical techniques, temperature measurement, charge measurement, capacity measurement, and more recently magnetoresistance. Also, atomic force microscopy has been used as a way of measuring interactions between molecules at the almost molecular level. The general structure of a biosensor is shown in Fig. 1.

Most of the efforts for the development of sensors and analytical methods has been for clinical diagnosis such as for glucose, and many of the technologies described here utilize applications to the clinical field. It should be recognized that there are many other areas and fields where sensors can now be used very effectively. This includes environmental analysis, fertilizers, toxic substances, the testing of foods, and also the testing of drugs, particularly hallucinogenic drugs. Further developments in sensor technology has allowed innovative methods for the measurement of DNA fragments. Recently discovered technologies of combinatorial chemistry allow the rapid selection of bioreceptors for specific analytes and also

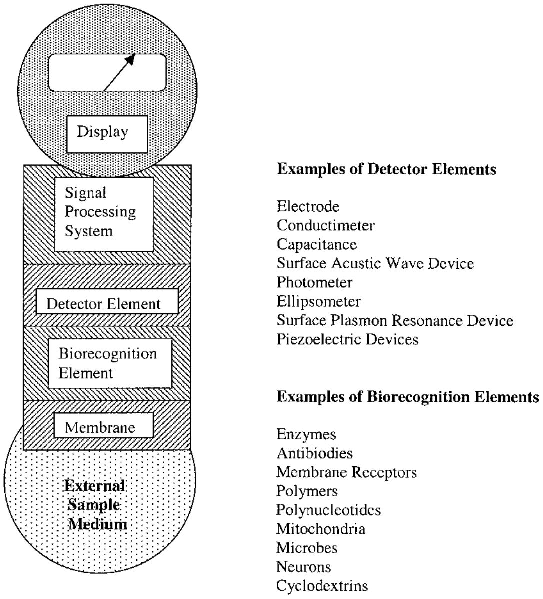


FIGURE 1 Key functional elements of a biosensor. Usually there is a membrane that separates the chemical and physical components of the sensor from the external environment containing the analyte of interest. The analyte diffuses through the membrane into the biochemical zone where it interacts with a biorecognition species to produce a change that is discernable by the detector element. A signal processing system then interprets this change (e.g., by comparing it to a calibration curve) to provide a readout of concentration.

with the correct sensitivity that is needed for a particular application.

I. MICROFABRICATION

Miniaturization of devices and mass production manufacturing techniques are some of the key reasons for commercial interest in biosensors at this time. Manufacturing technologies developed for completely different applications, such as micromachining for integrated circuits and fiber optics components for telecommunications, have allowed rather novel designs to be contemplated and developed for biosensors. Another key feature of these technologies is the miniaturization of devices that one is able to achieve. This capability leads to the potential of very small, sensitive, and very stable devices that allow the development of portable and perhaps disposable biosensors to permit bringing the analysis system very close to the source of the analyte rather the current mode of bringing samples of the analyte to centralized analysis laboratories. For example, if the application is environmental, one could put the sensor in remote locations to monitor many sites simultaneously. For medical uses one can put the sensor in a catheter to be inserted into a blood vessel, or an individual could have a portable analyzer and periodically place a drop of blood in contact with the sensor.

Usually the construction of a microsized analytical system has components for the following functions: sample injection, preparation, separation, and detection. The fabrication of these devices has been facilitated by the application of techniques that were originally developed for computing chip manufacture. While current techniques for computer devices can make features at the nanometer level, for microfabricated analytical devices the usual feature dimensions are more likely at the micron level.

The functions that are required for this devices are fluid handling systems: for fluid movement, combining liquid streams, splitting samples into multiple zones, reservoir zones for time delay and valves. There are several techniques for moving fluids in these micron-sized channels; these include capillary action, centrifugal force, gravity, and pump mechanisms. Figure 2 illustrates a peristaltic pump that has been microfabricated in a polymer substrate. The dimensions of the tubes are about 100 μ, and the rate of pumping as a function of the frequency of air pressure cycling is shown in the lower part of the figure.

One of the most effective techniques for fluid handling on microdevices utilizes electro-osmotic flow generation. This type of flow depends on generating an electrical double layer between the walls of the flow channel and the liquid in the channel. This mode of generating flow has one advantage because it generates a plug flow pattern, rather than a parabolic flow pattern that is typical of pressure-driven laminar flow. A plug flow pattern has minimum dispersion and thus limits the mixing between sample zones that can occur with laminar flow situations.

It was recently demonstrated that polynucleotide fragments can be separated in a microfluidic flow chamber that includes a series of "molecular dams." As illustrated in Fig. 3, large fragments of DNA are hindered by narrow passageways in the flow chamber. By appropriately deploying these barrier sections, a separation device can be constructed that can operate continuously.

Finally, electrophoresis is known as one of the most powerful separation techniques for the separation of biological materials. Capillary electrophoresis provides exceptionally high resolution of biomolecules. This technology has been microminiaturized as shown in Fig. 4. The resolution of a series of proteins is excellent and, very importantly, the reproducibility of different lanes is exceptional (Fig. 5).

An example of a commercial microanalysis system that utilized microfabrication and incorporated fluidic circuits

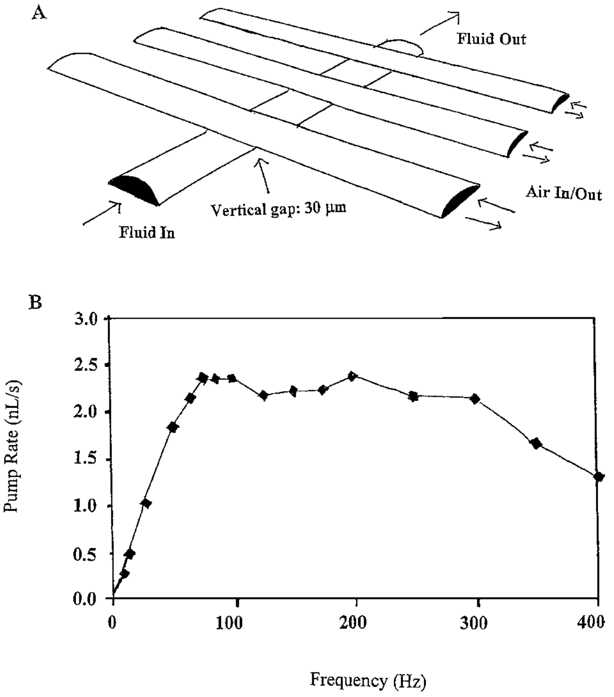


FIGURE 2 Microfluidics is a key technology for the chemical analysis of samples with microassay systems. This figure illustrates a peristaltic pumping technique that has been fabricated into a polymeric structure by soft lithography. (A) Channels are constructed by bonding layers of elastomers. Pulses of air pressure are sequentially applied to the upper "fingers" from bottom to top. This results in a directional squeezing of the lower channel to propel a liquid from one section of the device to another. (B) The effect of frequency on the pumping rate of fluid through the lower channel shows a linear response in the lower frequency range. [From Unger, M. A. *et al.* (2000). *Science* **288,** 113. Reprinted with permission of the American Association for the Advancement of Science, Washington, DC.]

and biosensors is the product produced by iSTAT for point-of-care monitoring of blood chemistries. An exploded view of their device is shown in Fig. 6. iSTAT gives the following explanation for the operation of their device:

The sensors are micro-fabricated thin film electrodes that are connected to amperometric, potentiometric or conductometric circuits. When blood samples contact the sensors, they are measured electrochemically as follows: Sodium, Potassium, Chloride, Ionized Calcium, pH and PCO_2 are measured by ion-selective electrode potentiometry. Concentrations are calculated from the measured potential through the Nernst equation. Urea is first hydrolyzed to ammonium ions in a reaction catalyzed by the enzyme urease. The ammonium ions are measured by an ion-selective electrode and concentration is calculated from the measured potential through the Nernst equation. Glucose is measured amperometrically. Oxidation of glucose, catalyzed by the enzyme glucose oxidase, produces hydrogen peroxide. The liberated hydrogen peroxide is oxidized at an electrode to produce an electric current which is proportional to the glucose concentration. PO_2 is measured amperometrically. The oxygen sensor is similar to a conventional Clark electrode. Oxygen permeates through a gas permeable membrane from the blood sample into an internal electrolyte solution where it is reduced at the cathode. The oxygen reduction current is proportional to the dissolved oxygen concentration. Hematocrit is determined conductometrically. The measured conductivity, after correction for electrolyte concentration, is related to the hematocrit. A variety of calculated results are available that include HCO^3, TCO_2, BE, sO_2, Anion Gap and Hemoglobin.

The concept of developing array technologies for massive analysis in biotechnology research was introduced in 1991. A group at Affymax showed that gene sequencing and gene discovery could be accomplished by immobilizing thousands of polynucleotide fragments of known characteristics as a matrix of micron-sized spots on a microscope-like slides. The array chip is then exposed to fluorescently labeled gene fragments from a sample under study. Hybridization is allowed to occur and then by the pattern of fluorescent spots on the chip the DNA sequence of the unknown sample can be deduced. Array technology has be expanded to be useful in identifying genetic defects and in drug discovery.

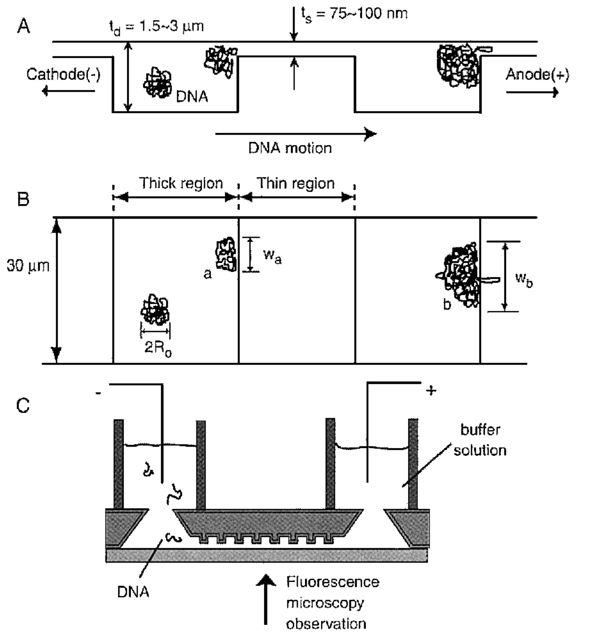


FIGURE 3 Microfabrication techniques allow a new method for the separation of macromolecules by electrophoresis. Molecules are driven across microchannel slits by an applied voltage. Mobility differences related to the size and structure of the macromolecule are created at these slits that result in a efficient separation of DNA samples. [From Han, J., and Craighead, H. G. (2000). *Science* **288,** 1026. Reprinted with permission of the American Association for the Advancement of Science, Washington, DC.]

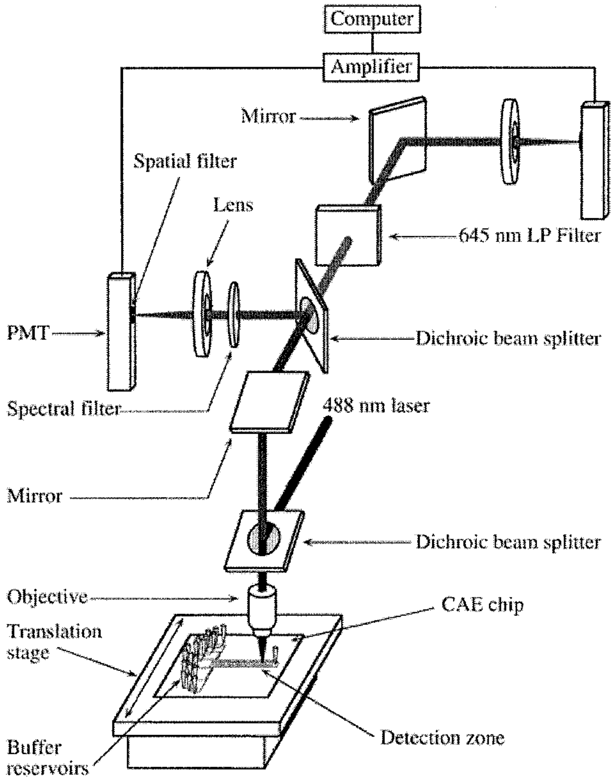


FIGURE 4 Capillary electrophoresis provides a very efficient and effective method to identify the components of complex mixtures. Microfabrication of multiple capillary channels on a chip allows rapid identification of mixtures because of the small distances involved, approximately 50 mm. This illustration shows the laser-excited confocal-fluorescence scanner that is used to determine the amounts of material passing the detection zone as a function of time. [From Woolley *et al.* (1977). *Anal. Chem.* **69,** 2183–2184. Reprinted with permission of the American Chemical Society, Washington, DC.]

II. BIOSENSORS

As illustrated in Fig. 1, there are three primary components of a biosensor: (1) the detector element, (2) the biological element, and (3) membranes used to separate the various structural elements of the sensor. The detector element performs the task of providing a signal related to the recognition event—that is, the result of the interaction of the analyte to be measured with the biological recognition molecule. The detector translates what is essentially a chemical interaction to some type of physical signal that can be manipulated by a variety of a electronic or optical techniques to finally produce an electrical output that is related to the amount of the analyte of interest.

The detector element is primarily responsible for the sensitivity of the device. Some of the very many detectors that have been tested are shown in Fig. 1. However, most of the detectors that have been used in commercial devices are electrochemical or optical in nature.

Perhaps the most unique component of a biosensor is the biological system that is utilized to identify specific molecules of interest in solutions of complex mixtures. The biological element of course is primarily responsible for the selectivity of biosensors. There are many different types of biological recognition systems that have been explored for sensors, ranging from the molecular scale—e.g., bioreceptors, enzymes, and antibodies—to cellular structures like mitochondria, and even immobilized whole cells and tissues. However, to date for practical reasons most commercially feasible biosensors have primarily utilized enzymes, and to a lesser extent antibodies.

Packaging is to some extent one of the most critical elements of a biosensor from the point of view of practicality. There have been hundreds of demonstrations of biosensor concepts using different types of biological recognition

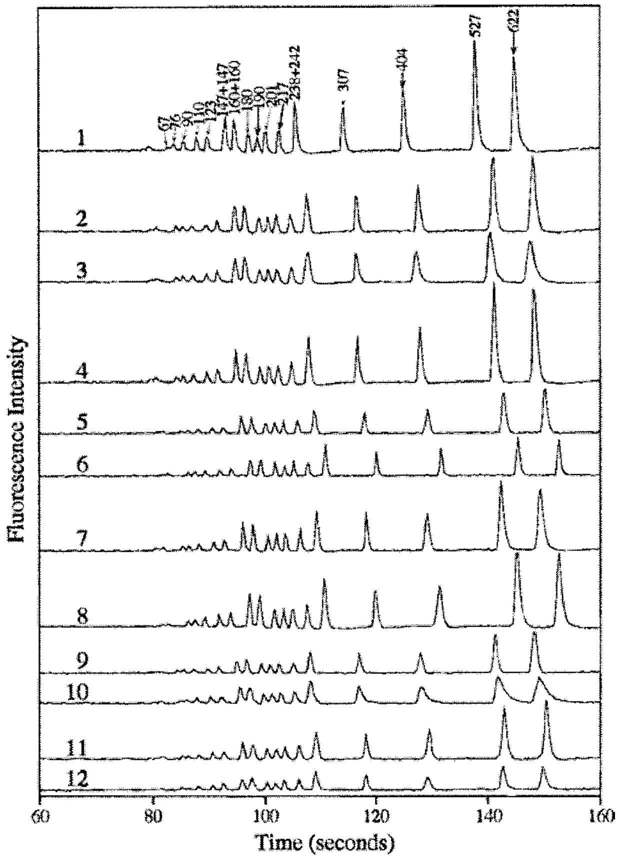


FIGURE 5 This figure illustrates that the entire analysis of mixture of polynucleotide fragments of from 67 to 622 base pairs can be accomplished in 160 sec. Also, the reproducibility from channel to channel allows the easy discrimination of minor changes in composition between samples. [From Woolley *et al.* (1997). *Anal. Chem.* **69,** 2183–2184. Reprinted with permission of the American Chemical Society, Washington, DC.]

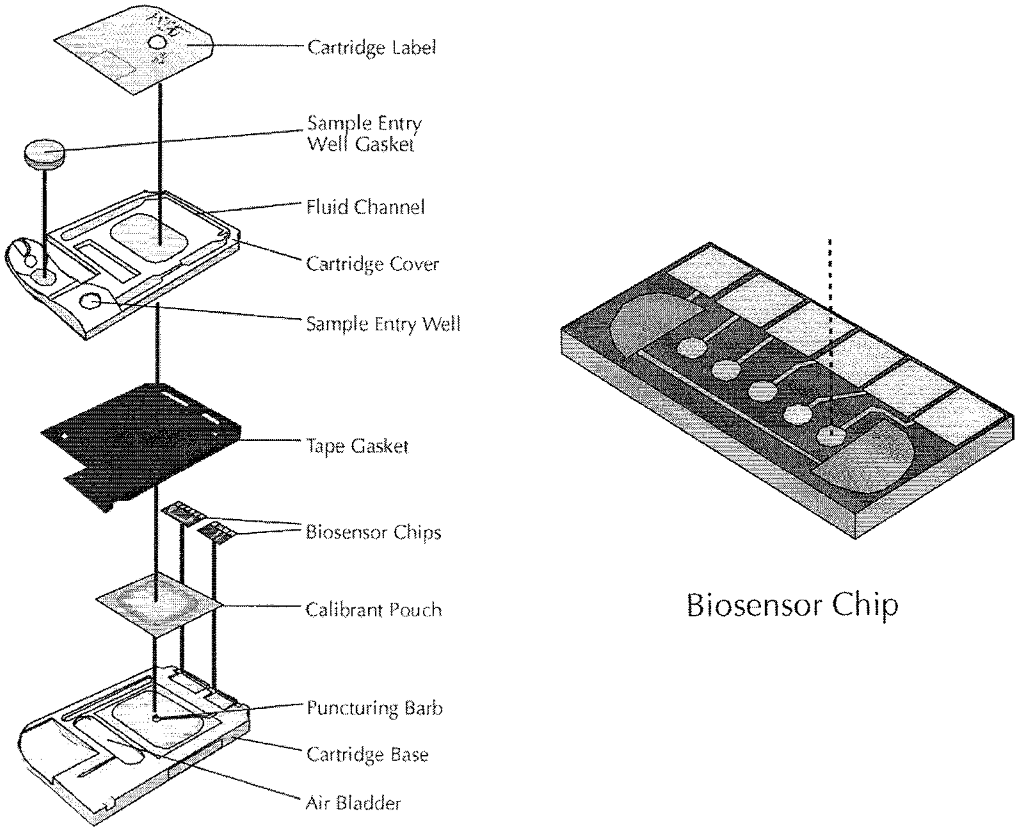


FIGURE 6 Exploded diagram of iSTAT's Portable Clinical Analyzer cartridge. This disposable cartridge measures approximately 1″ × 2″ and consists of calibration fluids as well as biosensors for the simultaneous analyses of up to eight different materials. After the blood sample is placed on the cartridge, it is inserted into a hand-held analyzer module that provides the controls for carrying out the analysis, display values for the analytes and a storage module for the information. An analysis is completed in minutes.

elements and detectors. However, a very limited number of all these possibilities have actually been translated into working biosensors because of the extreme difficulty in packaging of these devices to produce a product that is stable, that can be manufactured in large numbers, that is reproducible, and that can be used in many different environments.

Different types of membrane materials have been used in packaging sensors to provide some degree of selectivity, separation, and protection of the detector and biological elements of biosensors. And although a number of esoteric membrane preparations have been reported, for the most part, commercial biosensors utilize either (a) polymer films that are permeable to gases and exclude solutes such as salts and dissolved organics, or (b) microporous membranes are available with very large range of pore sizes—from reverse osmosis membranes that have pores on the order of 2–3 Å, to ultrafiltration membranes that may have pores on the order of hundreds of Å. These microporous membranes can provide selectivity for sensor application but they are a more difficult to manufacture integrally with the sensor. Dialysis membranes are permeable to low molecular weight solutes but retain or exclude macromolecules such as proteins. And then, finally, membranes or films can be "doped" with other reactive components to provide even higher levels of selectivity—for example, ion selective membranes.

One of the earliest biosensors was the glucose sensor devised by Leland Clark. (See Table I.) This prototype of biosensor is illustrated in Fig. 7, and the three primary components of the biosensor are identified. The biological recognition element, in this case glucose oxidase and catalase, are confined to a region between the oxygen sensor and the external sample by two membranes of different selectivity. The detector element is a polarographic electrode that measures the local oxygen concentration. The biorecognition element will be discussed first. There are a variety of substances that can be used for this purpose, but we can illustrate the concept with the use of an enzyme (Fig. 7). In this sensor the enzyme glucose oxidase was utilized to produce a change in composition that was proportional to glucose concentration. In the presence of

TABLE I Historical Landmarks in the Development of Chemical Sensors and Biosensors

Date	Event	Investigators
1916	First report on the immobilization of proteins: Adsorption of invertase to activated charcoal	J. M. Nelson and E. G. Griffin
1922	First glass pH electrode	W. S. Hughes
1954	Invention of the oxygen electrode	L. C. Clark, Jr.
	Invention of the pCO_2 electrode	R. W. Stow and B. F. Randall
1962	First amperometric biosensor: Glucose oxidase-based enzyme electrode for glucose	L. C. Clark, Jr., and C. Lyons
1964	Coated piezoelectric quartz crystals as sensors for water, hydrocarbons, polar molecules, and hydrogen sulfide	W. H. King, Jr.
1969	First potentiometric biosensor: Acrylamide-immobilized urease on an ammonia electrode to detect urea	G. Guilbault and J. Montalvo
1972–1974	First commercial enzyme electrode (for glucose) and glucose analyzer using the electrode (Yellow Springs Instruments)	
1975	First binding-protein biosensor: Immobilized concanavalin A in a polyvinyl chloride membrane on a platinum wire electrode to measure yeast mannan	J. Janata
	Invention of the pCO_2/pO_2 optrode	D. W. Lubbers and N. Opitz
1979	Surface acoustic wave sensors for gases	J. Wohltjen and R. Dessey
1979	Fiber-optic-based biosensor for glucose	J. S. Schultz and G. Sims
1980	Fiber-optic pH sensors for *in vivo* blood gases	J. I. Peterson, S. R. Goldstein, R. V. Fitzgerald, and D. K. Buckhold
1983	Molecular-level fabrication techniques and theory for molecular-level electronic devices	F. L. Carter
1986	First tissue-based biosensor: Antennules from blue crabs mounted in a chamber with a platinum electrode to detect amino acids	S. L. Belli and G. A. Rechnitz
1987	First receptor-based biosensor: Acetylcholine receptor on a capacitance transducer for cholinergics	R. F. Taylor, I. G. Marenchic, and E. J. Cook
1991	First array sensor on a chip	S. Fodor

glucose oxidase, glucose is oxidized to gluconic acid and hydrogen peroxide, which can further be decomposed to produce water with the net utilization of one mole of oxygen. This chemical reaction produces a number of changes such as pH, temperature, and H_2O_2 that could be detected for the purpose of creating a glucose sensor as discussed below.

$$\text{Glucose} + 2O_2 \rightarrow \text{Gluconic Acid} + H_2O_2,$$
$$2\,H_2O_2 \rightarrow 2\,H_2O + O_2.$$

However, Dr. Clark utilized the fact that one result of this catalyzed reaction is the depletion of the molecular oxygen in the vicinity of the enzyme, since the biological reaction utilizes one mole of oxygen for each mole of glucose that is oxidized. He showed that the enzymatically produced change in oxygen content can be detected polarographically by the so-called oxygen electrode, also developed by Dr. Clark. This particular device is a polarographic cell encapsulated by a gas-permeable polymer membrane such as a silicon rubber. Oxygen that diffuses into the cell reacts at a polarized electrode to form hydroxide, and the current is proportional to the oxygen concentration in the enzyme compartment. Clark's electrochemical cell works well because the platinum electrode is protected from the external environment by a polymer film; this film allows gases to pass relatively freely into the electrochemical compartment but prevents organic materials such as glucose or other electroactive substances from passing into the detector compartment. This combination of detector and enzyme biorecognition elements resulted in the first documented biosensor that became the basis for a series of commercial biosensors produced by the YSI Corporation, discussed below.

This sensor works in the presence other sugars, amino acids, or biochemicals in the sample fluid that might provide interference in other methods. The influence of these potential interferences on the output of the sensor will be minor because glucose oxidase has a very high selectivity for glucose and does not oxidize other compounds to any measurable extent.

One of the requirements of a practical biosensor is that the various components have to be assembled into a device that preserves the configuration of the various elements so that they maintain their functional capabilities over time. Figure 8 shows the structure of sensor meeting these requirements that is marketed by YSI; in this family of devices sample size is on the order of 10–25 μl. By using

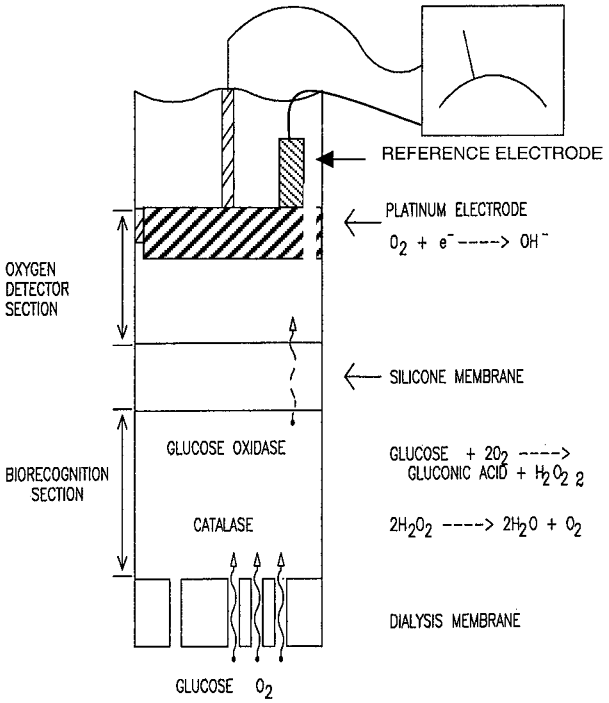


FIGURE 7 Components of the Clark glucose biosensor. This first portable biosensor is based on the measurement of the consumption of molecular oxygen due to the oxidation of glucose to gluconic acid as catalyzed by the enzyme glucose oxidase. Both glucose and oxygen diffuse into the sensor through a dialysis membrane. The first compartment encountered contains the enzymes that cause the oxidation of glucose with a simultaneous reduction in the ambient oxygen concentration. The second compartment contains a polarographic electrolytic oxygen measuring sensor that results in a current. Due to the consumption of oxygen in the first chamber, the current produced is inversely related to the glucose in the external fluid.

different immobilized enzyme membranes, YSI has produced a general sensor configuration to measure many different analytes. Some of the analytes that can be measured utilizing this enzyme biosensor are choline, D-glucose, dextrose, ethanol, galactose, hydrogen peroxide, lactose, L-glutamate, L-glutamine, L-lactic acid, methanol, and sucrose. This illustrates the range of selectivity that is obtainable using enzymes as the recognition element.

III. BIOLOGICAL RECOGNITION ELEMENTS

Figure 1 lists many other biological systems that could be utilized as the recognition element in biosensors. Generally speaking, one can select systems starting at the very simple molecular level and proceed through to much more complex systems, such as whole cells or even tissues. Here we limit discussion to three classes of biomolecules: enzymes, binding proteins (there are a number of binding substances that can be utilized in this fashion), and antibodies. Antibodies are large proteins (MW about 150,000) naturally found in blood, which are produced by the immune system for protection of an individual against foreign substances. Fortunately with new techniques in molecular biology, antibodies can be made to virtually any organic substance with a molecular weight greater than several hundred.

Enzymes are the class of bioactive agents that are the most readily adapted to biosensors because many enzymes are commercially available in large quantities and have been characterized in a great deal of detail. Antibodies would appear to be the most versatile type of biomolecule that can be employed in sensors, since they can be developed for virtually any analyte, but they are costly substances at this time. The most difficult kind biological substance to use would be tissues themselves because methods of sustaining stability are not well developed.

Enzymes are the choice for biosensors directed toward analytes that fall into the category of biochemicals that are metabolic intermediates. Because the metabolic enzymes (e.g., in glycolysis or protein synthesis) already exist in nature, and thus are a fruitful choice of materials that adapted for use in biosensors. Also, since there is vast technology associated with enzymes, usually one can find a commercial source for an enzyme for nearly any organic material. On the other hand, enzymes vary widely in their degree of specificity as compared with the relatively high

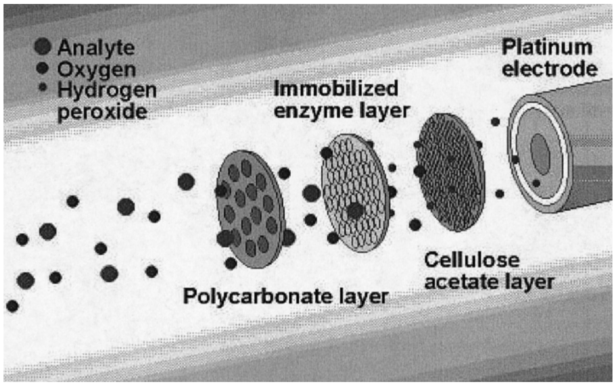


FIGURE 8 YSI Corporation manufactures a complete line of enzyme-based biosensors that follow from Clark's invention. By providing kits containing different enzymes immobilized on a membrane substrate, this generic approach allows a single detector device (a polarographic electrode) to be used for a number of analytes. In this system the electrode measures the hydrogen peroxide generated by the enzyme reaction.

TABLE II Enzymes That Have Been Used in Commercial Biosensors

		K_M (mol 1^{-1})	Substrate
Alcohol dehydrogenase	EC 1.1.3.13	1×10^{-2}	Ethanol
Catalase	EC 1.11.1.6	1	Sucrose, glucose
Choline dehydrogenase	EC 1.1.99.1	7×10^{-3}	Phospholipids
Lactate dehydrogenase	EC 1.1.2.3	2×10^{-3}	Lactate
Galactose oxidase	EC 1.1.3.9	2×10^{-1}	Galactose
Glycerol kinase	EC 2.7.1.30	1×10^{-6}	Glycerol
Glucose oxidase	EC 1.1.3.4	1×10^{-2}	Glucose
Lactate dehydrogenase	EC 1.1.1.27	2×10^{-2}	Lactate
Lactate oxidase	EC 1.1.3.2	2×10^{-2}	Lactate
Mutarotase	EC 5.1.3.3	3×10^{-2}	Glucose
Sulfite oxidase	EC 1.8.3.1	3×10^{-5}	Sulfite
Urate oxidase	EC 1.7.3.3	2×10^{-5}	Uric acid

degree of specificity of antibodies. Some of the common enzymes that have been used in sensor applications are shown in Table II.

Another class of biomolecules that can be used as recognition elements in sensors are bioreceptors. These are proteineous species that normally are found in the membranes of biological cells. Many of these proteins have been identified that have a role as signaling agents to provide information transfer between the two sides of a cell membrane—in other words, these molecules are the primitive manifestation of senses (taste, touch, smell, vision) for single cells. Membrane bioreceptors have been identified with a responsiveness to many different substances that either excite cells or to regulate the cell's behavior. Some examples are given below.

Receptor	*Candidate analyte*
Nicotinic	Organic phosphates
Adrenergic	Propanolol
Serotonergic	Imipramine
GABA	Barbiturates
Sodium channel	DDT
Estrogen	Tamoxifern
Auxin	2,4-D
E colicins	Aminobenzypurines

However, the number of bioreceptors that have been identified, numbering into the hundreds, is much more limited than the diversity of antibodies that number into the thousands; thus the number analytes that might be measured by biosensors based on bioreceptors is somewhat limited. On the other hand, this class of cell regulators that interact with bioreceptors may be very important biologically, e.g., hormones.

IV. IMMOBILIZATION

Technologies for immobilizing proteins on surfaces that have been developed over the last 25 years are particularly critical to the continued development of biosensors. There a dozens of techniques for activating surfaces so as to be able to covalently bond proteins while maintaining their biological activity. These methods have been used routinely for manufacturing adsorbents for affinity chromatography, dip-stick type analytical methods, e.g., immunoassays, and immobilized enzymes for commercial catalysts.

Techniques that are used to immobilize biological components for use in sensors include entrapment and encapsulation covalent binding crosslinking and absorption these various methods have been developed for proteins and there is a wide variety of chemical techniques to carry out each of these methods.

V. DETECTOR ELEMENTS

The appropriate detector element depends on the type reaction that is being affected by the enzyme. For example, for an enzyme that involves an oxidation-reduction mechanism, such as glucose oxidase, an electrochemical transducer would be a natural choice. If a substrate is utilized that produces a color, then one can use optical techniques to measure the extent of reaction, and if the enzymatic reaction results in a change in pH, then one can use a pH electrode to determine the extent of the reaction.

In addition to electrochemical and optical detectors, there have been a very wide variety of other kinds of detector systems that have been employed in experimental biosensors as noted in Fig. 1, including surface wave detectors, which are essentially mass detectors, conductivity, and thermal transducers. However, because of the need for reliability, flexibility, and sensitivity, electrochemical transducers are usually the system of choice because of the extensive research and manufacturing experience available.

The basic behavior of the electrochemical detector is shown in Fig. 9. Here the current response as a function of voltage applied is displayed. The rather nonlinear response is typical of electrochemical systems. Throughout the most of the range of applied voltages, positive or

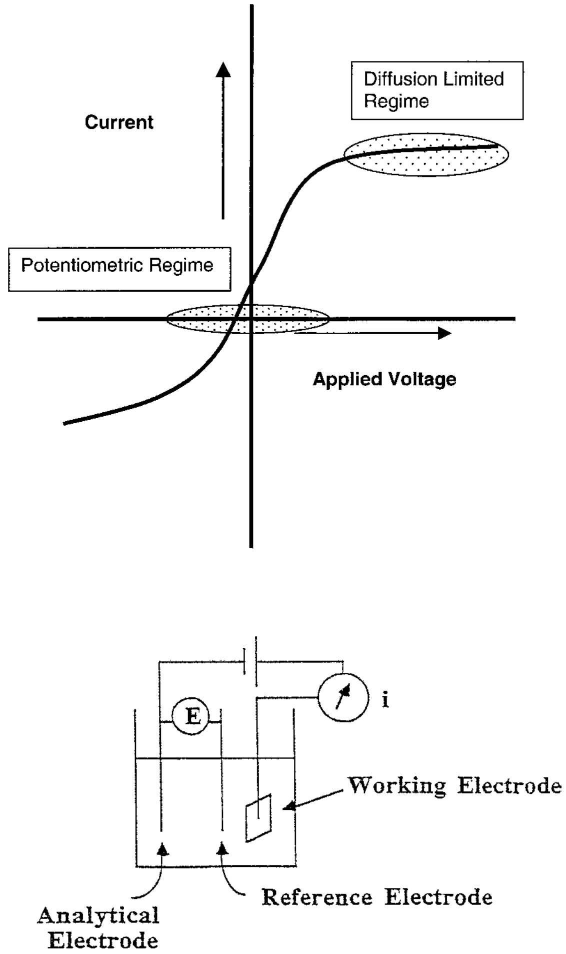


FIGURE 9 Schematic of the circuit components used in electroanalytical methods. The voltage applied to the inert analytical electrode (usually platinum) is show as E (usually between 0 and 1 volt), and the current measured (usually in microamps to milliamps) is shown as i. The lower graph shows the typical nonlinear response of current produced as a function of applied voltage. The magnitude of the current in the diffusion-limited regime is related to the concentration of the analyte.

negative currents flow through the electrode system, depending upon the magnitude and sign of applied voltage. In some ranges shown as shaded in the upper right, the current is rather independent of applied voltage. In these diffusion limited regions, the current is limited by the rate at which an electroactive species diffuses to the electrode surface and the magnitude of the current is proportional to the concentration of these species. These regions are called amperometric regions. There is one specific point when no current flows. The voltage under these conditions is related to the type of analyte reacting at the electrode and its concentration. Systems operating in this region, lower shaded region, are called potentiometric systems. By scanning voltages across the electrode, one can obtain both types of information. Usually in order to have reproducible behavior, one of the electrodes is made smaller than the other one, forcing the smaller electrode (working electrode) to determine the current voltage behavior.

An important difference in the operating characteristics of the two modes of operating electrochemical electrodes is that in the amperometric mode, the limiting current is directly proportional to the concentration of diffusion limited analyte. On the other hand, if the electrode is operated in the potentiometric region the output is related to the ratio of concentrations of oxidized and reduced forms of the analyte. Thus operating in potentiometric mode results in a system with a dynamic range of several orders of magnitude (on the order of several hundredfold in concentration), while the amperometric mode of operation usually has a dynamic range of tenfold.

Often in these devices one wants to be sure of the potential at the working electrode, and in these cases a three-electrode system is used that includes a reference electrode. By using a reference electrode, one is more certain of the electrical chemical events that are taking place at the working electrode. When an electrode is operated in an amperometric mode, the plateau regimes are known as film limiting current regions. This current depends on the rate of delivery of the electroactive species to the working electrode, which in the case of bare electrodes depends on the "stirring" in the vicinity of the electrode. One of the primary breakthroughs that set off the explosion in biosensors was Leland Clark's invention of placing a diffusion resistance film, in the form of a gas-permeable membrane, in front of the working electrode. This film had constant properties independent of the flow regime external to the sensor; this membrane then also determined the film thickness of liquid between the electrode and the membrane. Then the membrane covered electrode calibration was virtually independent of the mixing or turbulence in the sample environment.

One of the problems with using electrochemical methods for measuring concentrations is that if the electrode is presented with a mixture of substances, many which are electroactive, the measured current will be affected by the concentrations of all these various components. However in practice, a way to achieve a degree of stability is to assure that only one electroactive species reaches the region of the working electrode. This really means, then, that there must be some separation of the working electrode from the sample environment, which may contain many different materials. One way to accomplish this is by the use of membranes that are permeable to a very limited number of compounds. For example, the classical pH electrode utilizes a glass material that only allows protons to have any sufficient penetration or exchange, but other ions or organic materials virtually do not interact

and therefore do not affect the potential of the electrode inside the glass electrode. Thus the glass pH electrode is very robust device since it can be utilized in diverse environments.

One of the limitations of using electrodes as detectors is that many electroactive substances do not react at the electrode surface easily, causing variations in response with time. This problem can be overcome by utilizing highly reversibly electroactive substances (known as mediators) as intermediates between the analyte and electrode. Some common mediators are ferrocene, phenazine methosulfate, and benzylviologen. Since different organic materials have different potentials where they react in the electrode, is necessary to choose a mediator that has an oxidation/reduction potential that is near the oxidation/reduction potential of the analyte.

The typical behavior of a potentiometric based enzyme sensors is shown in Fig. 10 adopted from Rechnitz. In this design of a glutamine sensor, an organism was chosen that was particularly rich in glutamic deaminase. The reaction is

$$\text{Glutamine} \rightarrow \text{Glutamic Acid} + \text{Ammonia}$$

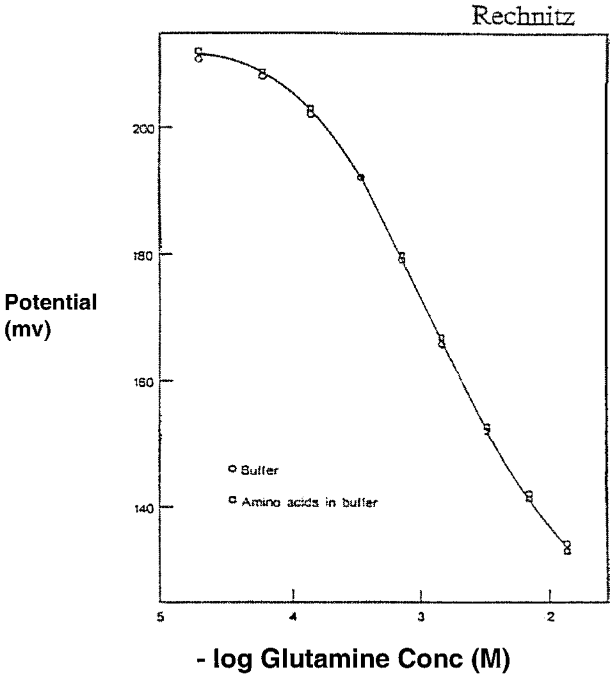


FIGURE 10 For ion-specific electrodes that operate in the potentiometer regime, concentration is related logarithmically to potential at the electrode surface. Dr. G. Rechnitz demonstrated the use of this enhanced response by developing a biosensor consisting of an enzyme layer containing glutamic deaminase interposed between an ion-selective ammonia electrode and the sample solution. The dynamic range of this type of biosensor is on the order of 100.

The extent of reaction, and therefore the glutamine concentration, was determined by measuring the ammonium content in the sensor. This was achieved with an ammonia electrode. Also shown is the response of the sensor in the presence of various potential interfering amino acids, which made a very minor contribution to the sensor response. This example illustrates the superb selectivity potential of enzymes due to their high degree of specificity. The response of the sensor is over three orders of magnitude of glutamine concentration, showing the wide dynamic range of potentiometric based transducers.

VI. OPTICALLY BASED BIOSENSORS

The use of optical methods became popularized after the work of Lubbers and Opitz, who developed technology that they termed optrodes. These are fiber-optic devices where there is chemically sensitive material placed at the terminal end of a fiber-optic system. Light is transmitted then along the optical fiber and when it interacts with the material at the terminal end of the fiber a change in the characteristics of the light occurs, e.g., absorption, fluorescence, scattering, polarization. Some of this light energy is captured and retransmitted back through the same or alternative fiber to a photodetector device. The change in optical characteristic can then be related to the concentration of material at the end of the optical fiber.

The structure of a typical fiber optic waveguide is shown in Fig. 11. These waveguides have primarily been developed for communication purposes and have the

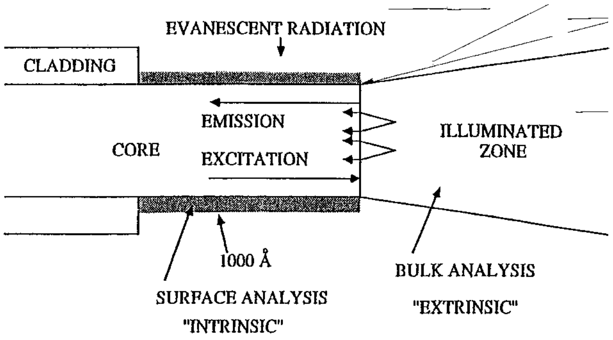


FIGURE 11 Fiber-optic waveguides are useful devices for biosensors based on optical techniques. Two types of illumination are available from optical fibers. If a fiber is cut, the terminal end provides a thin beam of light that can be directed into a sample region for optical determination of sample characteristics, typically the zone is about 0.1 mm in diameter and 2 mm in depth. If the cladding on the optically fiber is removed, then a zone of evanescent radiation is produced that illuminates a region up to 1000 Å from the glass surface. In this mode, chemicals that are further from the surface do not contribute significantly to the detected optical signal.

outstanding characteristic that the amount of light loss due to absorption and scattering is minuscule for path lengths on the order of meters. Also, the diameter of the fibers can be as small as 10 μ, although 150 μ is more usual. Thus, these optical waveguides can be used for transmitting optical signals almost without lost over the distances that one might need for biosensor monitoring devices. Hundreds of types of optical fibers are available in many different sizes and of materials, so that there is a large selection from which to choose.

There are essentially two modes by which light energy is dissipated from these waveguides, as shown in Fig. 11. First, light emanates from distal or cut end of the fiber, much in the same manner as the beam of light coming out of a flashlight. The intensity of this light is a function of distance from the end face and radial position off the center axis of the fiber. The magnitude of the intensity falls off fairly rapidly with distance away from the face of the optical fiber. Using this "illumination" zone for the measurement of analytes is usually referred to as using the endface configuration. Typically the intensity of light decreases with distance from the endface so that the practical optical path length is on the order of five or ten diameters of the optical fiber. For example, if the fiber is roughly 200 μ in diameter, then the practical path length of the illuminated zone is on the order of 1 ml.

There is another region of illumination that is of interest for the fabrication of optically based biosensors. If the cladding is removed from a portion of the optical fiber, there is a "light cloud" that hugs the surface, technically known as the evanescent radiation. One of the major properties of the evanescent radiation is that its intensity falls off much more rapidly with distance than light coming out of the endface. For practical applications, the sampling region for surface evanescent is on the order of a couple of hundred Å. Thus the evanescent mode is the method of choice for systems designed around measurement of surface-adsorbed antibodies. But the amount of light energy that is in the evanescent region is smaller on a unit area basis than the amount energy that is in the illuminating region at the distal end of a fiber, so that the amount of surface area that has to be exposed in order to get enough sensitivity the evanescent wave devices has to be much larger. Exposed lengths on the order of centimeters are required for this type of application.

Optical fiber detectors are usually utilized as ether miniature spectrophotometers or fluorimeters, because of the availability of indicator dyes that can be detected by absorbance or fluorescence methods. Figure 12 shows the absorption spectrum for phenol red, a dye commonly used as pH indicator. By measuring the absorption of solutions of phenol red at two wavelengths, the isobestic point where absorption is virtually independent of pH and at 600 mμ where the intensity of absorption is very pH sensitive, one can estimate the pH of a solution or region optically. Thus in the glutamine biosensor mentioned above, such a fiber-optic pH sensor could be substituted for the potentiometric ammonia sensor.

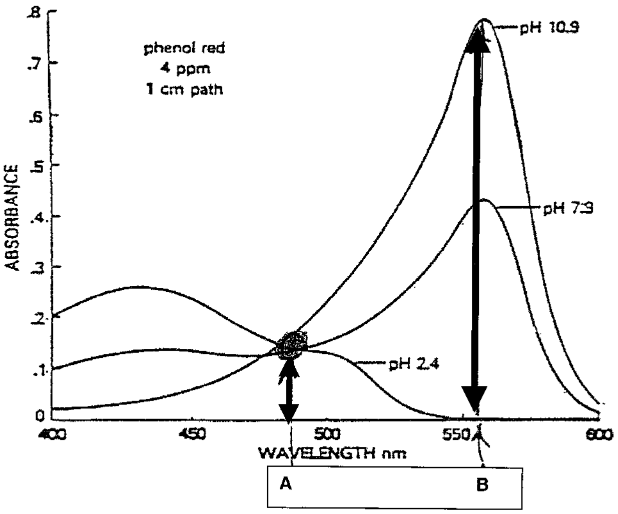


FIGURE 12 Concentrations of chemicals can be measured colorimetric ally. This example shows that pH can be estimated by measuring the absorption of light by a solution of phenol red at two wavelengths. At 550 nm the extinction coefficient of phenol red changes dramatically with pH, while at 470 nm the extinction coefficient is virtually independent of pH. By utilizing the ratio of absorbance at these two wavelengths, an accurate estimate of pH can be obtained.

Many different chromophores have been developed for the use in bioassays. These compounds are either changing color as a result of some change in the environment such as pH or they may change their fluorescent properties as a result of some chemical environmental change. Fluorescence methods tend to be more sensitive for analytical purposes if in the absence of the analyte light emission is minimal resulting in a desirable a high signal-to-noise ratio.

Characteristics of Some Fluorophores

Compound	*Excitation peak wavelength (nm)*	*Emission peak wavelength (nm)*
Lucifer Yellow	430	540
Fluorescein	492	520
Rhodamine B	550	585
Cy3.5	581	596
Cy5	649	670
Cy5.5	675	694
Cy7	743	767
Nd-benzoyltrifluoroacetone	800	900

One of the important features of fiber optic technology is that optical multiplexers are available to connect many sensors to the same spectrophotometer or fluorimeter.

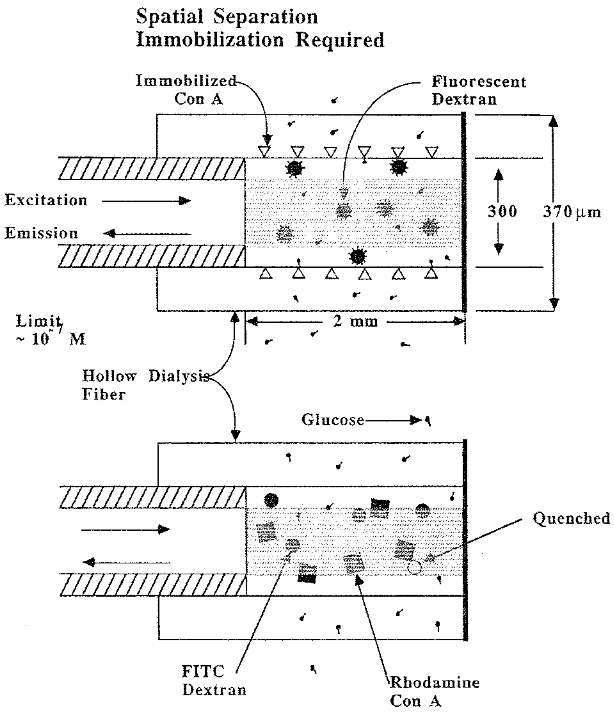


FIGURE 13 Dr. Jerome Schultz adapted the principles of immunoassays to make biosensor devices (sometimes called immunosensors or affinity sensors). In this example Concanavalin A serves as the surrogate antibody for glucose and fluorescently labeled dextran as the analog to the antigen. In the top figure ConA is immobilized in the interior surface of a hollow dialysis fiber. In the absence of glucose, most of the dextran binds to the ConA out of the field of view of the optical fiber. In the presence of glucose the dextran is displaced from the ConA, moves into the illuminated zone, and produces a fluorescent signal that is picked up by the optical fiber. In the lower figure an alternative strategy (called Fluorescent Energy Transfer or FRET) is used to measure the competitive binding of dextran and glucose for ConA sites. Here the ConA is not immobilized by labeled with Rhodamine. In the absence of glucose the potential fluorescence from dextran that binds to ConA is quenched due to the close proximity of Rhodamine. In the presence of glucose, Dextran is displaced from ConA and its fluorescence is detected by the optical fiber.

Figure 13 shows a biosensor based on fiber optics developed by Jerome Schultz that illustrates some of the principles of bioreceptor-based sensors and fiber optics. The approach is to use a biospecific macromolecule that has binding selectivity for the analyte of interest. To make a glucose sensor Concanavalin A (Con A), a lectin that has selective binding sites for carbohydrates was chosen as the biorecognition agent. This methodology is similar to that used in immunoassays: it is based on the competition between the analyte (glucose) and an analog of the analyte (Dextran), which has a fluorescent label. A difference between this approach and immunoassays is that the binding interactions are reversible and there is no need to replenish the reagents. A hollow-fiber dialysis membrane is used to form a microscopic porous test tube, and the Con A is immobilized on the interior surface membrane. This ensures retention of both the Dextran and Con A in the sensor. In the absence of glucose in the external ‘ medium, most of the FITC Dextran will be occupying sites on the Con A out of the view of the light that comes out of the distal end of the optical fiber, and thus there is very little of fluorescence that enters back into the optical fiber. On the other hand, if the responsive end of the optic fiber is placed in a solution containing sugar, glucose can diffuse through the wall of the dialysis tubing and compete for Con A binding sites, displacing some of the FITC Dextran that then distributes uniformly throughout the lumen of the hollow fiber, and the fraction in the middle of the hollow fiber is excited by the light that comes out of the optical fiber. A portion of the emitted fluorescence from FITC Dextran is captured by the same optical fiber producing, transmitted to a photomultiplier that produces a signal directly related to the glucose concentration in the external medium.

A typical calibration curve is shown in Fig. 14. The dynamic range of this type of sensor is less than potentiometric methods because the receptor sites become saturated at high concentrations of analyte. In contrast to immunoassays, the sensor response is reversible.

Another important optical technique is surface plasmon resonance, a phenomenon that has been known for a long time but has recently been applied for measuring large biomolecules. Figure 15 shows the configuration for plasmon surface resonance sensors. Basically the principal of these devices is based on the fact that when a protein is absorbed on the surface it changes the index of refraction of that surface. When the surface is illuminated at various angles, then the angle at which one gets a maximum reflection is a function of the amount of material absorbed at the surface. These devices can have a high degree of sensitivity; however, they also may have problems

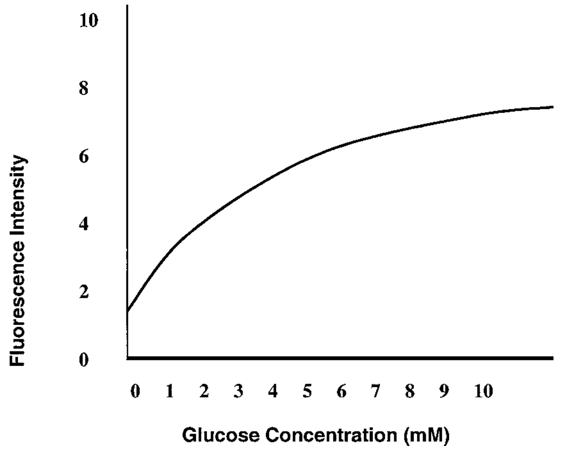


FIGURE 14 A typical calibration curve for an affinity sensor shows a leveling-in response at high analyte levels because all the analog analyte is displaced from the bioreceptor.

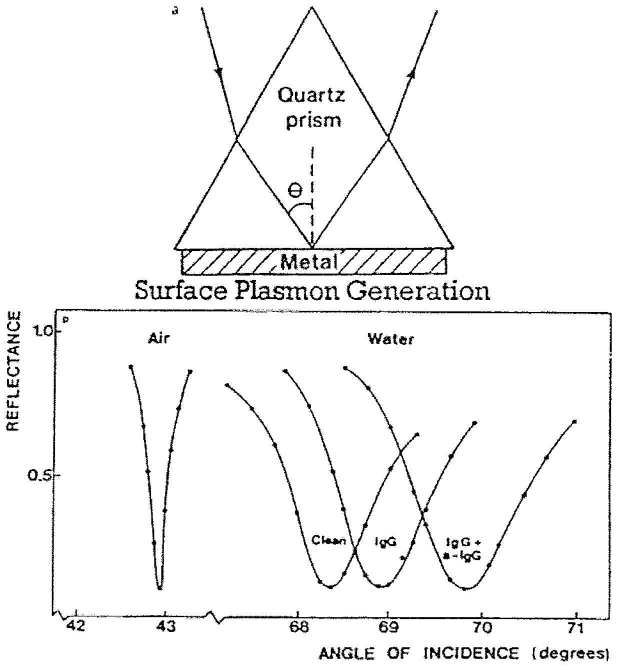


FIGURE 15 The principle of biosensors based on surface plasmon resonance. The angle of total internal reflection is a function of the refractive index change at the interface between the metalized surface and the sample medium. The lower figure shows that a large change the angle of maximum reflection occurs between air and water. And in water a minor but measurable effect is seen when an antibody is adsorbed on the surface.

in that the selectivity could be low if materials not related to the analyte can indiscriminately absorb to the surface.

Another principal that has been evaluated for the construction of optically based sensors is the use of chemiluminescence. In these cases an enzyme system specific for the analyte are coupled to reactions that produce light through chemiluminescence. In principle, systems of this type could be very sensitive, first of all due to the amplification factor of enzyme reactions, and secondarily because fluorescence measurements are among the most sensitive of optical techniques.

VII. OTHER DETECTORS

One can also measure a temperature produced by the chemical reaction if the systems is quite well insolated and temperature differences are small as 10^{-6} °C can be detected with sensitive bridge techniques and this can measure materials as low as 10^{-5} molar.

Another type of detection device is based on oscillating quartz crystals. These types of devices are very similar to the systems that are used inside of electronic watches. The limit of detection with these systems where there is a frequency change based on the amount of material that attaches to the surface can be as low as 10^{-12} g. Surface acoustic wave detectors can be used as well.

VIII. SENSOR DYNAMICS

The response time of sensors to changes in concentration of an analyte depends on the response rate of the various components of a sensor. Usually, diffusional processes are the limiting factor; these include diffusion of the analyte to the surface of the sensor (dependent on external mixing), diffusion through membranes, and diffusion through the various regions of the sensor structure. Since diffusion lag times increase with the square of distance, it is imperative to maintain the active layers of the sensor to dimensions on the order of tenths of a millimeter or less. Enzyme reactions can usually be made nonlimiting if necessary by increasing the amount of enzyme in the system. However, the rate of dissociation of antibody–analyte complexes decreases directly with increasing binding affinity. For high levels of sensitivity, high binding constants are needed. For example, to measure an analyte at a concentration of nanomoles, a binding constant of the order of nanomoles is required. The dissociation rate for such antibodies is on the order of tens of minutes.

Some of the considerations that are important in determining the structure of sensors for use in analytical assays include the following:

Sensitivity: Sensitivity relates to the lowest concentration that the system can reliably detect. Table III shows that the concentration levels must be able to measure varies quite greatly among different biochemicals that are important in the body. The biochemicals in blood can have average concentration ranges from milligrams per cubic centimeter to nanograms per cubic centimeter. The wide range of concentration of different biochemicals that are present in blood presents a technical challenge to the measurement of several analytes simultaneously in the same detection system.

Dynamic range: In addition to sensitivity of an assay system, the range of sensitivity of the device is another important consideration. The dynamic range is usually defined as the ratio of highest concentration to the lowest concentration that a particular technique can reliably measure. For many analytical methods based on biosensors, the dynamic range is rarely more than a factor of about 10. Thus for any particular application, the system needs to be engineered to achieve dynamic range that covers the normal variation in concentration for that analyte. For example, for the measurement of blood glucose the normal level is about 100 mg/dl and the dynamic range should be between 50 and 200 mg/dl and somewhat higher for diabetics. For systems that use

TABLE III Approximate Blood Levels of Plasma Constituents

Compound	$\times 10^{-3}$ M	Compound	$\times 10^{-6}$ M
Sodium	140	Albumin	600
Chloride	105	Urate	250
Bicarbonate	30	Phenylalanine	125
Glucose	5	Immunoglobulin G	90
Urea	4	Ammonia	30
Cholesterol	4	Iron	20
Calcium	2.5	Bilirubin (Total)	10
Triglycerides (fasting)	1	Immunoglobulin M	1
Compound	**$\times 10^{-9}$ M**	**Compound**	**$\times 10^{-12}$ M**
Oestriol (late pregnancy)	600	Aldosterone	180
Thyroxine binding globulin	500	Insulin	120
Cortisol	400	Parathyroid hormone	100
Placental lactogen	300	Growth hormone	50
Thyroxine (total)	125	Luteinising hormone	10
Corticosterone	20	Triiodothyronine (free)	10
Triiodothyronine	2	Adrenocorticotrophin	10
Prolactin	1	Thyroid stimulating hormone	5
Oestradiol-17 (women)	1	Angiotensin II	4
Progesterone	1	Oxytocin	1
		Arginine vasopressin	1

biological materials such as enzymes or antibodies, this means that one has to select the appropriate protein with the correct binding characteristics or reactive characteristics for the molecule that is to be measured. Fortunately, with the advent of modern molecular biology, one can either select or modify the protein to fulfill the analytical requirements. An equation that depicts the interaction of an antibody with a ligand is

$$\text{Antibody} + \text{Ligand} \underset{k_r}{\overset{k_f}{\rightleftarrows}} \text{Antibody} - \text{Ligand}.$$

The affinity constant for these antibiodies is the concentration at which the protein is half saturated with the ligand.

$$K_a = k_f/k_r.$$

Enzymes have a kinetic characteristics that show saturation behavior:

$$v = V_m\, SE/(K_M + S).$$

The parameter for enzymes that characterizes the sensitivity level for sensors based on the uses of enzymes is the kinetic parameter known as the Michaelis–Menton constant (K_M). Again enzymes can be found in nature or readily genetically modified to have K_M that vary over 5 orders of magnitude. Table II lists K_M values for some enzymes that have been used in biosensors.

Response time: The response behavior of sensor-based analytical devices is affected by a number of parameters. With respect to the sensor component, response characteristics are usually determined by a combination of diffusional and reaction kinetic processes. The usual configuration for a biosensor includes a membrane that protects the sensor element. Thus in the Clark glucose sensor shown in Figure 7, the diffusion processes include the passage of glucose and oxygen from the external fluid through the dialysis membrane into the enzyme compartment and the subsequent passage of oxygen through the silicon rubber membrane into the electrode compartment.

The approximate lag time introduced by diffusion layers (either membranes or liquid layers) can be expressed as

$$\text{Lag time (sec)} = 0.2L^2/D,$$

where D is the diffusivity of solute in the layer (cm^2/sec) and L is the thickness of layer (cm).

If the biological element is an enzyme, then the kinetics of the reaction of analyte will be governed by an equation similar to that given above, and the overall rate of reaction will be determined by the enzyme concentration in the preparation (E) and the inherent substrate turnover of the enzyme characterized by V_m.

If the biological element is an antibody or a bioreceptor, the kinetics will be related to the magnitude of the affinity constant K_a. The affinity constant is essentially the equilibrium constant for the reversible association and dissociation of the ligand for the antibody. Kinetic studies of these reactions has shown that increase in binding constants is directly correlated with a reduction in the inherent dissociation rates of these reactions. The implication of this behavior is that the response rate of sensors based on antibodies becomes slower with higher affinities.

An approximate value for the association rate constant (k_f) for ligands with antibodies is 10^8 M^{-1} sec^{-1}. Thus an approximation for the lag time for dissociation of antibiody–ligand complexes is

$$\text{Lag time (sec)} = 10^{-8}/K_a,$$

where K_a is the ligand–antibody association equilibrium constant (M^{-1}).

SEE ALSO THE FOLLOWING ARTICLES

BIOENERGETICS • BIOMATERIALS, SYNTHESIS, FABRICATION, AND APPLICATIONS • BIOPOLYMERS • ELECTRO-

PHORESIS • ENZYME MECHANISMS • IMMUNOLOGY–AUTOIMMUNITY • ION TRANSPORT ACROSS BIOLOGICAL MEMBRANES • MACROMOLECULES, STRUCTURE • METABOLIC ENGINEERING

BIBLIOGRAPHY

Buck, R. P., *et al.* (1990). "Biosensor Technology. Fundamentals and Applications," Harold Dekker, New York. 419 pp.

Cass, A. (1990). "Biosensors. A Practical Approach," IRL Press at Oxford University Press, Oxford.

Hall, A. H. (1990). "Biosensors," Open University Press, Buckingham.

Janata, I. (1989). "Principles of Chemical Sensors," Plenum Press, New York.

Kress-Rogers, E. (1996). "Handbook of Biosensors and Electronic Noses," CRC Press, New York.

Taylor, R. F., and Schultz, J. S. (1996). "Handbook of Chemical and Biological Sensors," Institute of Physics Press, Philadelphia.

Turner, A. P. F., Karube, I., and Wilson, G. S. (1987). "Biosensors. Fundamentals and Applications," Oxford Science Publications, Oxford.

Optical Fiber Techniques for Medical Applications

Abraham Katzir
Tel Aviv University

I. Introduction
II. Optical Fibers
III. Lasers for Fiberoptic Medical Systems
IV. Fiberoptic Endoscopes
V. Fiberoptic Medical Diagnostics
VI. Integrated Fiberoptic Systems
VII. Laser–Fiberoptic Systems and Their Clinical Applications
VIII. Novel Fiberoptic Medical Systems
IX. Outlook

GLOSSARY

Acceptance angle Maximum incident angle for which an optical fiber will transmit light by total internal reflection.

Catheter Flexible hollow tube normally employed to inject liquids into or to drain fluids from body cavities.

Cladding Outer part of an optical fiber; has a lower refractive index than the core.

Core Inner part of an optical fiber; has a higher refractive index than the cladding layer.

Critical angle Minimum incidence angle in a medium of higher refractive index for which light is totally internally reflected.

Endoscope Optical instrument used for viewing internal organs.

Fiberoptic endoscope A flexible endoscope that contains a fiberscope and ancillary channels for medical instruments and for irrigation or suction.

Fiberscope Viewing instrument that incorporates an ordered bundle for imaging and a fiber bundle for illumination.

Laser catheter Catheter that incorporates an optical fiber for the transmission of a laser beam.

Laser endoscope Endoscope that incorporates an optical fiber or rod lens system for the transmission of a laser beam.

Light guide Assembly of optical fibers that are bundled but not ordered; used for illumination.

Numerical aperture Light-gathering power of an optical fiber. It is proportional to the sine of the acceptance angle.

Optical fiber Thin thread of transparent material through which light can be transmitted by total internal reflection.

Ordered bundle Assembly of optical fibers in which the fibers are ordered in exactly the same way at both ends of the bundle.

Power fiber Optical fiber that can transmit a laser beam of high intensity.

Rigid endoscope Sometime called "rigid telescope." Imaging in this device is carried out by a series of lenses that are placed inside a rigid tube.

Total internal reflection Reflection of light at the interface between media of different refractive indices, when the angle of incidence is larger than a critical angle.

Videoscope A special endoscope that contains a miniature electronic imaging device (e.g., charge-coupled device) instead of an ordered imaging fiber bundle.

OPTICAL FIBERS are thin, flexible, and transparent guides through which light can be transmitted from one end to another. An ordered bundle of optical fibers forms a fiberscope and can be used for transmitting images. A medical endoscope incorporates such a fiberscope. Thin, flexible endoscopes enable physicians to get images from areas inside the body that were previously inaccessible, such as inside the bronchial tree or inside blood vessels or even inside the fetus in the womb. Optical fibers can be used for medical diagnostics by inserting them (e.g., through catheters) inside the body and making physical or chemical measurements through them. Measurements such as blood flow, pressure, and gas content or sugar content in blood can be performed quickly and reliably. Optical power fibers may be used to transmit high laser power to areas inside the body. The laser radiation in turn may be used to cut tissues, coagulate bleeding vessels, remove tumors, clean occluded arteries, and destroy cancer cells. Compound laser endoscopes may include several channels: fiberoptic ones for image transmission, diagnostics, and laser power transmission, and ancillary channels for injection of liquids or aspiration of debris. Laser endoscopes are thin and flexible and used in myriad applications.

I. INTRODUCTION

Light can be transmitted through a cylinder of transparent material by a series of internal reflections. This phenomenon was probably known by ancient glass blowers, but the earliest demonstration of the effect was given by J. Tyndall in England in 1870. Thin rods that are used for light transmission are called optical fibers. Such fibers appear in nature. For example, tissues of plant seedlings can guide light to coordinate their physiology.

An ordered bundle of glass or plastic fibers may be used for the transmission of images. Some of the earliest papers and patents describing optical fibers, in the period 1925–1930, involved the use of such bundles in medicine. Only in the early 1950s, with the development of optical fibers that consisted of inner core and outer cladding, did the flexible fiberoptic bundle become a powerful practical tool. At first the medical instruments, endoscopes, incorporating fiberoptic bundles, enabled physicians to obtain images from organs inside the body.

During the 1970s and 1980s there has been rapid progress in the development of glass fibers with extremely low optical transmission losses. The main challenge in the use of these fibers is in communication systems. Light signals emitted from semiconductor lasers can be transmitted through very long optical fibers and can be detected by semiconductor detectors. The laser light can be modulated in such a way that it may carry with it voice communication, radio and television broadcasting, and computer data. Thus, optical fibers will be the major building blocks in future communication networks.

There has been increasing recognition of the enormous possibilities for utilizing the low-loss optical fibers in medical applications. Ultrathin endoscopes with improved optical properties have been developed and utilized for endoscopic imaging. Optical fibers have been used as diagnostic tools. Laser power has been transmitted through optical fibers and used for ablation of tissue or for therapeutic applications inside the body. The diagnostic fibers and the "power" fibers can be incorporated in compound laser endoscopes. The unique capabilities of this system may be enhanced by computer control. This will actually make it a robotic system that will interpret the imaging and the diagnosis data, and will control the therapeutic or the surgical laser operation inside the body. We review here some of the physical principles involved and some of the revolutionary uses of optical fibers in medicine.

II. OPTICAL FIBERS

In this section we describe the physical principles of light guiding in optical fibers and in bundles of fibers. We discuss the materials and methods of fabrication of various fibers and bundles.

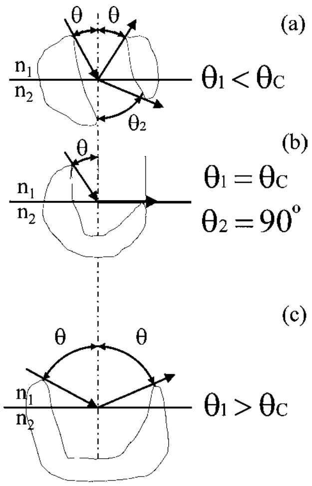


FIGURE 1 Reflection and refraction at the interface between two media of refractive indices n_1 and n_2 ($n_1 > n_2$). (a) Incident angle $\theta_1 < \theta_c$. (b) Incident angle $\theta_1 = \theta_c$. (c) Incident angle $\theta_1 > \theta_c$; total internal reflection.

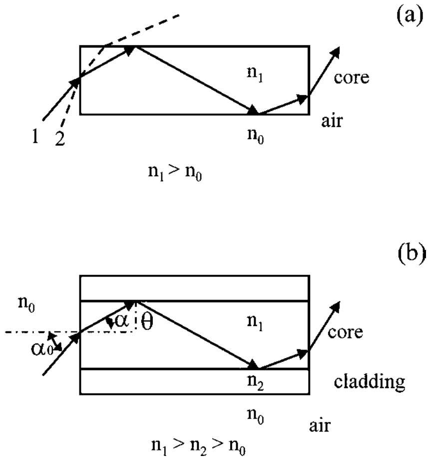


FIGURE 2 Trajectory of a ray in a cylindrical rod. (a) A rod of refractive index n_1 in air ($n_0 = 1$). (b) A rod whose core has a refractive index n_1 and cladding index n_2 ($n_2 < n_1$).

A. Properties of Optical Fibers

1. Total Internal Reflection

Consider two transparent media of refractive indices n_1 and n_2, where $n_1 > n_2$ (Fig. 1). A ray of light propagates at an angle θ_1 with respect to the normal to the interface between the media. At the interface, part of the beam will be reflected back to medium 1, and part will be refracted into medium 2. The reflection is specular (the angle of reflection is equal to the angle of incidence θ_1). The refraction obeys Snell's law as shown in Fig. 1(a), so that $n_1 \sin\theta_1 = n_2 \sin\theta_2$. If the angle θ_1 is increased, one may reach some angle $\theta_1 = \theta_{1c}$, for which $\theta_2 = 90°$. Here θ_{1c} is called the critical angle, and for this angle we write $n_1 \sin \theta_{1c} = n_2 \sin 90° = n_2$. For every angle of incidence $\theta_1 > \theta_{1c}$, there is no refracted beam. Were there such a beam, its angle θ_2 would be given by the equation: $\sin \theta_2 = (n_1 \sin\theta_1)/n_2 > (n_1 \sin\theta_{1c})/n_2 = 1$. Since this of course is impossible, there is only a reflected beam. This phenomenon, shown in Fig. 1(c), is called total internal reflection. If medium 1 is glass with $n_1 = 1.5$ and medium 2 is air with $n_2 = 1$, the critical angle is given by $\sin \theta_{1c} = 1/1.5$. In this case, $\theta_{1c} = 42°$. If medium 2 is soda-lime glass with $n_2 = 1.52$ and medium 1 is flint glass with $n_1 = 1.67$, then $\theta_{1c} = 65°$. It should be mentioned that in practice the total internal reflection is very efficient. In this process more than 0.99999 of the incident energy is reflected, compared to about 0.95 for good metal mirrors.

2. Optical Fibers

Consider a rod of transparent material of refractive index n_1 in air ($n_0 = 1$). Two rays of light incident on one end face of the rod are shown in Fig. 2(a). Ray II will be refracted inside the rod and refracted back in air. Ray I, on the other hand, will be totally internally reflected inside the rod and will emerge from the second face of the rod. This will also happen when the rod is very thin and flexible, and in this case the rod is called an optical fiber. A fiber consisting of a transparent material in air is called an unclad fiber. Light will propagate inside the rod (or fiber) by a series of internal reflections, even if the rod is not in air, as long as $n_2 < n_1$. In particular, a compound rod may consist of the inner part of index n_1, called core, and the outer part of index n_2, called cladding. If this rod is thin, the optical fiber formed is called a clad fiber. The cross section of the rod (or the fiber) is shown in Fig. 2(b), with the trajectory of an incident ray of light. Assume that the angle of incidence in air is α_0 and that inside the core the beam is refracted at an angle α_1, as shown. We can write $n_0 \sin\alpha_0 = n_1 \sin\alpha_1 = n_1 \cos\theta_1 = n_1(1 - \sin^2\theta_1)^{1/2}$, where $n_0 = 1$ is the refractive index in air. The angle θ_1 can assume sev eral values, but its maximum value for total internal reflection is the critical value θ_{1c}, given by $n_1 \sin\theta_{1c} = -n_2$. We can calculate $\alpha_{0\max}$, the value of α_0 corresponding to this value of θ: $n_0 \sin\alpha_{0\max} = n_1(1 - \sin^2 \theta_{1c})^{1/2} = (n_1^2 - n_2^2)^{1/2}$. This value of $n_0 \sin\alpha_{0\max}$ is defined as the numerical aperture NA: $\mathrm{NA} = n_0 \sin \alpha_{0\max} = (n_1^2 - n_2^2)^{1/2} \cdot (2\Delta)^{1/2}$, where $\Delta = (n_1 - n_2)/n_1$. Rays of light impinging on the surface at any angle $\alpha_0 < \alpha_{0\max}$ will be transmitted through the rod (or the optical fiber). All these rays form a cone of angle $\alpha_{0\max}$. For angles of incidence $\alpha_0 > \alpha_{0\max}$, the ray will be refracted into the cladding and then into the air. The cone of light rays that could be transmitted by the fiber is a measure for the light-gathering capability of the fiber. The angle of the

cone $\alpha_{0\max}$ is called the acceptance angle. If $n_1 = 1.62$ and $n_2 = 1.52$, then NA = 0.56. If the fiber is in air, $n_0 = 1$ and $\alpha_{0\max} = 34°$, but if it is immersed in water, $n_0 = 1.3$ and then $\alpha_{0\max} = 25°$. The acceptance angle of the fiber in water is therefore smaller than in air.

In the case of a straight fiber, a ray incident at an angle α will be transmitted by the fiber and emerge with the same angle α. For an incident cone of rays, the light emerging from the fiber will also be a cone with the same apex angle. In general, this is also true for a bent fiber.

3. Transmission in Optical Fibers

A beam of light of intensity I_0 impinges on a transparent material of thickness L (cm). The beam is transmitted through the material and emerges with intensity I. In many cases the ratio between I_0 and I is given by the Beer–Lambert law, $I = I_0 \exp(-\delta L)$, where δ (cm^{-1}) is the absorption coefficient. Physically, after traveling $L = 1/\delta$ cm, the intensity I_0 is reduced to $I_0/e \approx I_0/3$, and after traveling $L = 2/\delta$ cm it is reduced to $I_0/e^2 \approx I_0/10$. In the engineering literature, the transmission loss A is given in decibels (dB), as defined by $A\,(dB) = 10 \log_{10} (I_0/I)$. As an example, if only 10% of the light is transmitted, then $I = I_0/10$ and the transmission loss is $A = 10 \log_{10} 10 = 10$ dB. If the light intensity is reduced to I after traveling a distance L in the material, the transmission losses are stated in A/L dB/m (or dB/km). If the intensity is reduced from I_0 to $I_0/10$ after traveling in a fiber of length 10 meters, then the transmission loss is 10 dB/10 meters = 1 dB/meter.

A beam of light of wavelength λ and of intensity I_0 may propagate in a clad optical fiber, as shown in Fig. 3. The total transmission through real fibers depends on many factors, and several of them are listed here:

1. Losses in the coupling of light from a light source into the fiber.
2. Reflection losses at each end of the fiber. This is called the Fresnel Reflection and is proportional to $[(n_1 - n_0)/(n_1 + n_0)]^2$.
3. Absorption in the pure material of the core. This absorption depends on λ.
4. Absorption by impurities.
5. Scattering from the inherent inhomogeneous structure of the core (Rayleigh scattering); scattering from small inhomogeneity of refractive index in the core, and in the interface between core and cladding.
6. If the fiber is bent, some of the light leaks out from the cladding.

In order to ensure high transmission through the fiber, one has to reduce the losses. In unclad fibers the outer surface of the fiber is exposed to mechanical or chemical damage, which gives rise to scattering. Therefore the fiber has to be a clad fiber. The cladding material should be well matched to the core one, and the interface should be of high quality. The core should be very pure, in order to reduce absorption, be free of defects, and homogeneous, to reduce scattering. The ray trajectory shown in Fig. 3 is schematic. Some of the optical power is transmitted through the cladding. Therefore the cladding layer should also be highly transparent. In medical applications, low losses are very important when trying to transmit high optical power through optical fibers. High losses may give rise to excessive heating and to melting of the power-carrying fiber.

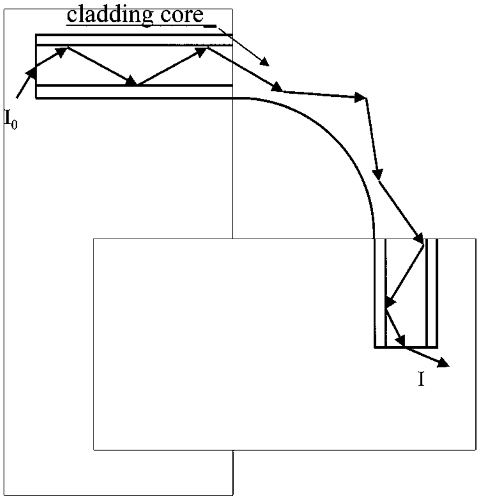


FIGURE 3 The transmission of light in a clad fiber.

B. Material and Methods of Fabrication of Optical Fibers

1. Optical Fibers Made of Silica Glass

Optical fibers are usually made of glass, and most often this is an oxide glass based on silica (SiO_2) and some additives. A rod of glass (called preform) is heated until soft, and then a fiber is drawn from its end. This fiber consists only of a core. The preform itself may consist of two glasses, of different indices of refraction, such as shown schematically in Fig. 2(b). In this case a core/clad fiber is fabricated by drawing. In both cases the outer diameter of the fiber is less than 0.1 mm, in order to obtain flexibility. Fibers in use today are generally core/clad fibers.

In the past the preform itself was prepared by melting, and the glass components were not very pure; also,

there were defects in the materials and at the interface. Nowadays, exceptionally pure materials are prepared by deposition in a vacuum system, and special efforts are taken to reduce the number of defects. In the case of optical communications, the fibers are coated with a plastic primary coating as they are being drawn, in order to protect them from moisture or scratches, among other things. Good-quality optical fibers have transmission losses on the order of 1 dB/km. Such quality is normally not needed for medical applications.

In optics, wavelength λ is normally specified in nanometers ($1\,nm = 10^{-9}m$), or in micrometers ($1\,\mu m = 10^{-6}$ m). The optical spectrum is broadly divided into three spectral regions: ultraviolet (UV) for $\lambda < 400$ nm, visible (VIS) $\lambda = 400$–700 nm, and infrared (lR) for $\lambda > 700$ nm. The UV region at wavelengths $\lambda < 250$ nm is sometimes called "deep UV." The infrared spectrum may be further divided to near infrared (NIR) $\lambda = 0.7$–3.0 μm and middle infrared (MIR) $\lambda = 3.0$–30 μm. Optical fibers that are based on some mixture of SiO_2, B_2O_3, and Na_2O, are called silica fibers and they transmit well in the UV, visible, and NIR regions of the optical spectra.

2. Special Optical Fibers

Standard silica fibers transmit only in the visible and do not transmit in the deep UV or in the middle infrared. Nor could these glass fibers transmit visible or NIR beams of high intensity. For the transmission of deep UV, one could use pure silica as the core material. Doped silica (of lower index of refraction) serves as a cladding layer. These fibers (sometime called "quartz" fibers) also serve as power fibers for the transmission of laser power. For transmission in the middle infrared, one may use nonoxide glasses such as chalcogenides (e.g., $As_2\,S_3$) or fluorides (e.g., ZrF_4–BaF_2–LaF_3). Crystals of silver halides or thallium halide can be extruded through dies. The resulting fibers were found to be transparent in the middle infrared. An alternative is to use hollow plastic or glass tubes, coated on the inside with thin metallic coatings. The hollow tubes are transparent in the mid IR and can deliver high laser power. These are not actually fibers, but they are thin and fairly flexible and are sometimes referred to as hollow fibers.

Optical fibers made of plastic materials have been used in the past for the visible range. For example, polymethylmethacrylate (PMMA) could be drawn to form unclad fibers. Alternatively, polystyrene may serve as a core in these fibers, and PMMA as cladding. Such fibers are flexible and less expensive to fabricate, but their properties are still inferior to those of glass. In particular, plastic fibers are relatively thick (1 mm), they transmit well only in the spectral range $\lambda > 450$ nm, and their laser power handling is limited.

III. LASERS FOR FIBEROPTIC MEDICAL SYSTEMS

A. Medical Lasers

There are many regular light sources that are used in medicine, such as incandescent lamps (e.g., halogen), high-pressure lamps (e.g., Xe), and light-emitting diodes (i.e., LEDs). The light emitted from these sources consists of many wavelengths; it is emitted from a relatively broad area, and is spread in many directions. Laser light has characteristics that make it especially useful in medicine. These are:

- Monochromatic—laser light consists of a narrow band of wavelengths (i.e., "one wavelength").
- Collimated beam—the laser light is emitted in a parallel beam.
- High intensity—practically all the laser energy is concentrated in a narrow pencil.
- Coherent—the laser emission is "ordered" both in space and in time.

The coherence helps to focus the beam to a very small point (of the order of the wavelength of light). The energy density (energy divided by area) at the focal spot is extremely high. Such energy density cannot be obtained with standard (noncoherent) light sources.

All lasers are based on some active medium in which "optical amplification" occurs. This medium can be a solid crystal (e.g., Nd:YAG or Er:YAG crystals), a liquid (e.g., dye in a solution), gas (e.g., Ar, CO_2, or HeNe), or a semiconductor (e.g., GaAs). The medium has to be excited (pumped) in order to start the lasing process, and electric current often provides excitation. Also, in order to get laser emission the laser should contain two mirrors, which reflect the light back and forth through the lasing medium. One of these mirrors is partially transparent, and the laser beam emerges from this mirror.

There are many lasers that emit light in the UV, visible, and IR. Some lasers emit continuously light of wavelength λ, and the total power emitted is measured in watts. These are called CW (continuous wave) lasers. Other lasers emit short bursts of light, in which case the lasers are called pulsed lasers. Some of the lasers used in medicine emit long pulses, measured in msec (10^{-3} sec), others emit shorter pulses, measured in μsec (10^{-6} sec) or nsec(10^{-9} sec), and some emit very short pulses, measured in psec (10^{-12} sec) or fsec (10^{-15} sec). The energy emitted from these lasers can be specified by the pulse

TABLE Ia Continuous Wave Lasers

Laser	Wavelength (μm)	Laser medium	Max power (W)
Ar ion	0.488 and 0.514	Gas	20
Dye	0.4–1.0	Liquid	2
HeNe	0.628	Gas	0.05
GaAs/GaAlAs	0.7–1.5	Semiconductor	>100
Nd:YAG	1.06	Solid	>100
CO_2	10.6	Gas	>100

TABLE Ib Pulsed Laser

Laser	Wavelength (μm)	Laser medium	Pulse duration	Max repetition rate	Max average power
Excimer	0.193, 0.249, 0.308, 0.351	Gas	5–30 nsec	50–150 Hz	10 W
Dye	0.4–1.0	Liquid	0.01–100 μsec	1000	5
Nd:YAG	1.06	Solid	0.01–100 μsec	1–100	>100
Ho:YAG	2.1	Solid	1–100 μsec	20	2
Er:YAG	2.94	Solid	5–200 μsec	10	20
CO_2	10.6	Gas	0.01–100 μsec	1000	>100

length (seconds), number of pulses per second (pps), energy per pulse (joules), and average power (watts). Some of the important lasers, which have been used in medicine, are given in Tables Ia and Ib. The typical values refer to lasers that are used for medical applications.

B. The Use of Various Lasers in Medicine

As a laser beam impinges on biological tissue, its energy is attenuated in the tissue, due to scattering and to absorption. In the absorption process, the energy is transformed to another form of energy—often to heat. The Beer–Lambert law gives the attenuation due to absorption in general. For every tissue and every wavelength λ there is an absorption coefficient a_λ, and one could then write $I_\lambda = I_\lambda(0)\exp(-xa_\lambda)$, for a tissue thickness x. For wavelength λ, and for $L_e = 4.6/a_\lambda$ (cm), $I_\lambda(L_e) = I_\lambda(0)\exp[-a_\lambda(4.6/a_\lambda)] = 0.01\, I_\lambda(0)$, so that 99% of the light is absorbed after traveling $L_e = 4.6/a_\lambda$ cm in tissue. L_e is the extinction length.

Absorption of laser light in tissues and in biological fluids depends on the characteristic absorption of water, hemoglobin, melanin, keratin, and protein. Soft tissues contain a very high percentage of water. Water does not absorb in the visible, but the absorption is high in the deep UV and the middle IR ($\lambda > 2.8\,\mu$m). For the CO_2 laser ($\lambda = 10.6\,\mu$m), the extinction length L_e is less than 0.1 mm. The same is true for the CO laser ($\lambda = 5\,\mu$m) and for Er:YAG ($\lambda = 2.94\,\mu$m). Red pigment in hemoglobin absorbs blue-green light, and therefore the extinction length of argon or krypton laser light in blood is also small. On the other hand, Nd:YAG ($\lambda = 1.06\,\mu$m) or GaAs ($\lambda = 0.8\,\mu$m) laser light is not absorbed well by tissue, and the extinction length L_e is a few millimeters. In the UV, excimer laser light ($\lambda < 300$ nm) is highly absorbed by both hard tissues and soft tissues, and L_e is small.

In the visible or in the infrared, the absorbed laser energy is often converted into heat. For low laser incidence (incident energy per unit area), tissue is moderately heated. This may give rise to therapeutic heating (such as laser induced thermal therapy) or to the coagulation of blood. For high incidence, the tissue vaporizes. In most cases it is water that boils away. In practice the laser beam is focused on a certain area, and the vaporization removes tissue from this area. The absorption of high intensity laser pulses may give rise to mechanical effects, such as the generation of shock waves. These may be used for the removal of hard tissues, such as gallbladder or urinary stones. The common medical lasers are bulky and heavy. The output laser beam is often delivered to a desired spot via an articulating arm. This is a system of mirrors, attached to a distal hand piece, which includes a focusing lens. This articulating arm enables the user to focus the beam and to move it from place to place. By moving the beam the physician can ablate tissue or perform an incision or an excision (e.g., laser surgery). Tissues outside the area of the focal spot are also affected. They heat up to a temperature that is dependent on the thermal diffusivity and the extinction length of the tissue. The heating effect may cause damage to surrounding tissue. The damage is lower if one uses

lasers for which the extinction length is small and if one uses short pulses, rather than a CW beam. The interaction described above is basically photothermal. The interaction between excimer lasers ($\lambda < 300$ nm) and tissue may be different. It involves the absorption of UV light, and there are claims that ablation of tissue is possibly due to photochemical processes that do not generate heat. The excimer laser has important surgical applications. It has been used in cosmetic surgery for resurfacing of the skin, such as for the removal of wrinkles, and in ophthalmology for corneal reshaping (i.e., for the correction of vision).

There are several nonthermal effects which should also be mentioned. The interaction of tissues with low level laser light (e.g., blue light or UV light) may give rise to luminescence. This luminescence can be used for tissue diagnosis, for example, for distinguishing between healthy and diseased tissues. Ar lasers or some new semiconductor lasers, which emit in the blue, can be used for these applications. Another important application is photochemotherapy. Chemicals are introduced into the body and are "triggered" by light of some specific wavelength and then selectively interact with tissue. Such photochemical effects may be used for cancer treatment, to be discussed in Section VII, or for the treatment of age–related degenerate macula (ADM) and other diseases.

Medical laser systems have dramatically improved during the last few years. Semiconductor lasers cover a broader range from the blue (e.g., GaN lasers) to the mid–IR (i.e., quantum cascade lasers). Solid-state lasers pumped by semiconductor lasers are now more compact and more powerful. There are gas lasers (e.g., CO_2) that are very compact and easy to use. All these new lasers are now more reliable and more efficient, most of them are lightweight, easy to operate, and relatively inexpensive. Many of the new lasers have already made their debut in the clinical setting.

C. Lasers and Fibers

Laser beams in the visible and NIR ($\lambda = 0.3$–$3.0\ \mu$m) are transmitted by silica-glass fibers and have been used for transmitting the radiation of Ar, dye, Nd:YAG, and GaAs lasers. Excimer laser radiation at $\lambda = 250$–300 nm can be transmitted only through pure silica fibers, which are not very useful at shorter wavelengths. Better fibers are needed for this spectral range. Er:YAG laser radiation can be transmitted through sapphire fibers. There has been a major effort to find suitable fibers for the transmission of the radiation of mid-infrared lasers, and especially that of a CO_2 laser ($\lambda = 10.6\ \mu$m). At the moment, the best fibers are hollow fibers or polycrystalline fibers made of halide crystals (e.g., AgClBr). The transmission of high laser power through fibers is discussed in Section VII.

IV. FIBEROPTIC ENDOSCOPES

The name endoscope is based on two Greek words: endon (within) and skopein (view). It is used to describe optical instruments that facilitate visual inspection and photography of internal organs. Endoscopes may be inserted into the body through natural openings (ear, throat, rectum, etc.) or through a small incision in the skin. The simplest endoscope is an open tube, and it has been used in medicine for thousands of years. The development of modern endoscopes started about a hundred years ago with the addition of artificial illumination (i.e., incandescent lamp) and lenses to the open tube (NITZE 1894). Rigid endoscopes built in a similar manner are still in use. The nature of endoscopy changed dramatically in the early 1950s when optical-fiber bundles were introduced. These are used both for illumination and for imaging.

A. Light Guides for Illumination

1. Light Sources

In endoscopy light is transmitted through an optical fiber, in order to illuminate an internal organ. The light sources that are suitable for this purpose should deliver enough energy through the fiber to facilitate viewing or photography. In general, high-intensity light sources are needed, such as tungsten lamps, mercury or xenon high-pressure arc lamps, or quartz iodine lamps. In most of these cases, the light has to be focused on a fiber (or a fiber bundle), by means of a lens or a reflector. Special provisions are often required to dissipate the excessive heat generated by the light sources. Some of the lasers that were mentioned earlier are suitable for special endoscopic illumination.

2. Light Guides (Nonordered Bundles)

The light from a regular high-intensity (noncoherent) source cannot be focused to a spot whose size is equal to the size of a thin optical fiber. Therefore, a bundle of fibers is normally used for illumination. This assembly consists of numerous clad fibers of a certain length, which are bundled together but not ordered. In order to increase the light collection efficiency, the fibers are designed to have a relatively high numerical aperture (NA 0.65–0.72). The diameter of each individual fiber is 20–50 μm, so it is flexible. The ends of the fibers are cemented with epoxy resin or fused. The remaining lengths are left free, so that the whole bundle can be bent. A thin plastic sleeve is normally used to protect the fibers and often a metal tube protects the ends of the bundle. It is also important that the NA of the collecting fibers match the NA of the optical element that focuses light from the light source onto the light guide.

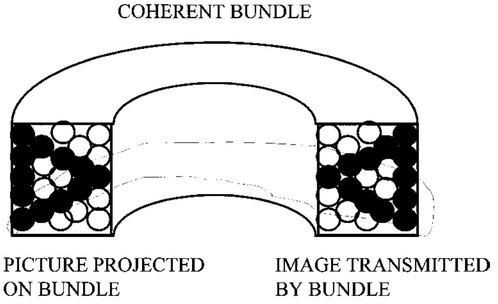


FIGURE 4 Image transmission through a coherent (aligned) bundle.

B. Ordered Bundles for Image Transmission

Optical fibers can be accurately aligned in a bundle, such that the order of the fibers at one end is identical to the order in the other end. Such an ordered optical-fiber assembly is sometimes called a coherent bundle (not to be confused with the coherent nature of laser light). If a picture is projected and imaged on one end of such a bundle, each individual fiber transmits the light impinging on it, and the ordered array will transmit the picture to the other end. This is shown schematically in Fig. 4.

The individual fibers used in a bundle are clad fibers. If unclad fibers were used, light would leak from a fiber to its nearest neighbors (crosstalk), and the quality of the image would deteriorate. The optical fibers in a bundle (or at least on each end) are closely packed, as shown schematically in Fig. 5. We define d as diameter of the cladding and c as the diameter of the core.

An ordered bundle has to faithfully transmit the image of an illuminated object. In order to get a bright image, as required for photography, each individual fiber should have a high NA. One should judiciously choose the materials needed for fabricating core and cladding. The thickness of the cladding layer of each individual fiber $(d - c)/2$ should be of the order of 1.5–2.5 μm, to ensure guiding with little loss and to minimize crosstalk. On the other hand, in order to get high spatial resolution one has to decrease the diameter c of the fiber core to 5–10 μm. In this case, the ratio between the areas of the cores to that of the bundle is not high, and the light transmission through the bundle is reduced. The fibers should therefore have low loss in order to obtain a good picture.

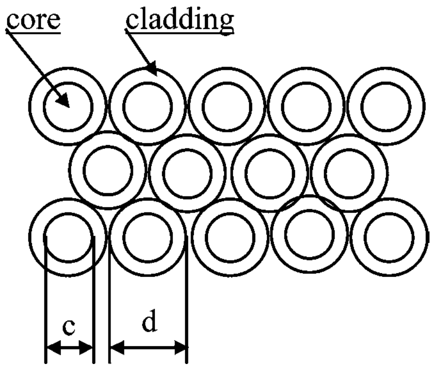


FIGURE 5 Close packing of clad optical fibers in a bundle, with *c* the diameter of the core and *d* the diameter of the cladding.

C. Fabrication of Fiberoptic Bundles

An ordered bundle is fabricated by winding a clad optical fiber on a precision machine and continuously winding many layers, one on top of another. All the fibers are then cut, at one point, glued together at each end, and polished. Tens of thousands of fibers, each of diameter of tens of micrometers, can be bundled in this way. Alternatively, one starts with an assembly of carefully aligned lengths of clad fibers. The assembly is heated in a furnace, and a compound fiber is drawn—much like the drawing of a single fiber. The resulting fiber is called a multifiber and may contain fibers of diameter of about 5 μm. If the number of fibers in a multifiber is small (e.g., 100), it is quite flexible. A few multifibers can then be aligned to form a flexible ordered bundle of very high spatial resolution. If the number of fibers is higher than 1000, the multifiber is rigid. It is then often called an image conduit and is used for image transmission in an ultrathin rigid endoscope ($d < 2$ mm).

Finally, a flexible multifiber may be fabricated by a leaching process. In this case, a special clad fiber is fabricated with two cladding layers. The inner cladding layer is made of an acid-resistant glass and has a lower refractive index than the core. The outer cladding layer is made of glass that is soluble in acid. The multifiber is made in the same way described above. It is then cut to a desired length and a plastic material is applied to protect the two ends. The multifiber is then immersed in an acid bath, and the outer cladding layer of the individual fibers is leached out and are separated from each other. A flexible ordered (aligned) bundle may thus consist of tens of thousands of fibers, each of diameter 10–50 μm.

D. Fiberscopes and Endoscopes

With the availability of fiberoptic bundles, one is able to design and fabricate a viewing instrument called a fiberscope. The structure of a rigid or a flexible fiberscope is shown schematically in Fig. 6.

Light from a lamp (or a laser) is focused onto the input end of a flexible light guide. The light is transmitted through this nonordered bundle and illuminates an object. An objective lens forms an image of the object on the distal end face of an ordered bundle. The image is transmitted through the bundle to the proximal end and viewed through

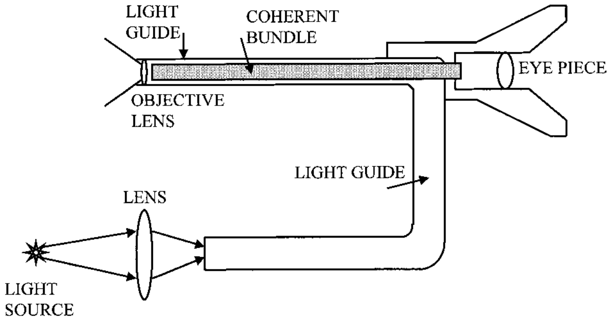


FIGURE 6 Schematic diagram of a fiberscope.

an eyepiece. The development of modern digital electronics has let to the development of miniature imaging devices called charge-couple devices (CCDs). These are used in digital (still) cameras or in video cameras. CCDs offer imaging on a TV monitor as well as easy data transmission and storage. A standard camera (incorporating photographic film) or a CCDs could be attached to the proximal end of the imaging bundle, for imaging applications.

Flexible endoscopes incorporate fiberscopes that enable the physician to view and examine organs and tissues inside the body. In addition, an endoscope may include ancillary channels through which the physician can perform other tasks. A schematic cross section of an endoscope is shown in Fig. 7. One open channel serves for inserting biopsy forceps, snares, and other instruments. Another channel may serve for insufflation with air or for injection of transparent liquids to clear the blood away and to improve visualization. A separate channel may also serve for aspiration or suction of liquids. Many endoscopes have some means for flexing the distal tip to facilitate better viewing.

The quality of fiberoptic endoscopes has markedly increased during the last few years. Several ultrathin endoscopes (see Section IV.E.2) have been tried clinically

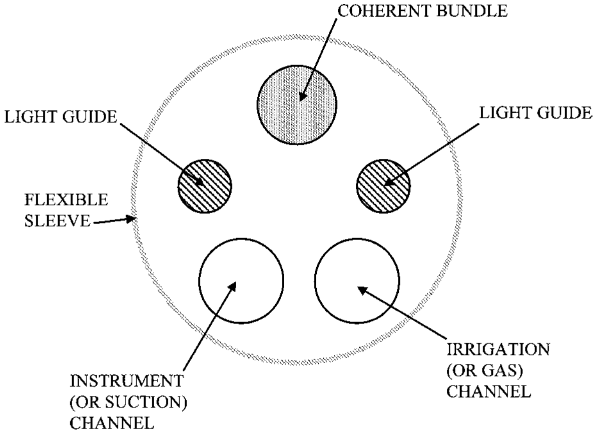


FIGURE 7 Schematic cross section of an endoscope.

TABLE II Typical Data on Commercially Available Flexible Endoscopes

Length	300–2500 mm
Outer diameter	0.5–15 mm
Instrumental channel diameter	0.5–3 mm
Flexible up/down	180°/60°
Depth of focus	5–100 mm
Field of view	50°–100°

in cardiology or ENT. There have also been attempts to develop disposable endoscopes. Rigid disposable endoscopes may incorporate a small number of glass or plastic lenses, and the flexible disposable endoscopes may incorporate plastic fibers. One of the limitations of standard endoscopes is that the image is a two-dimensional one. There are attempts to develop rigid endoscopes that incorporate two miniature imaging devices or fiberoptic endoscopes that incorporate two bundles of ordered bundles. Such endoscopes could be used for three-dimensional (3D) imaging (Table II). Research is being carried out in new areas such as developing specialized computer programs for image processing.

Some of the CCDs mentioned above are a few millimeters in diameter. Such a CCD could be placed at the distal end of an endoscope and used for imaging, instead of the ordered bundle of optical fibers. A light guide is still used in these endoscopes for illumination. Such endoscopes can provide imaging at a video rate from internal organs and they are called videoscopes. It is estimated that roughly 50% of the flexible endoscopes used today are videoscopes. These endoscopes are still relatively thick and they are mostly used in gastroscopy and colonoscopy.

E. Clinical Applications of Endoscopes

There are many types of endoscopes. Some of the endoscopes are flexible and others are rigid. Some are inserted through natural orifices, and some are inserted via a rigid tube through an incision in the skin (percutaneously). All these endoscopes are similar, from the basic science point of view. But, many of the mechanical and optical details vary according to the specific application. Fiberoptic endoscopes have been used in a variety of medical applications.

1. Standard Endoscopes

Standard fiberoptic endoscopes make use of an imaging bundle and one or several illumination bundles. Imaging through these endoscopes with white light illumination or with laser light can be used for early detection of diseases such as cancer. All these endoscopes incorporate an

ancillary channel through which surgical instruments can be inserted and used for endoscopic therapy or surgery or for the removal of tissues, such as a biopsy or the removal of tumors.

Some endoscopes are listed in alphabetical order:

Arthroscope—for the joints
Bronchoscope—for the bronchi
Colonoscope—for the colon
Colposcope—for the cervix and the vagina
Cystoscope—for the urinary bladder
Gastroscope—for the stomach, esophagus, and the bile duct
Hysteroscope—for the uterine cavity
Laparoscope—for the abdominal cavity
Laryngoscope—for the larynx
Otoscope—for the ear
Sinuscope—for the nose
Thoracoscope—for the thorax.

Some of the applications of these endoscopes are listed below:

The first fiberoptic endoscopes—the **Gastroscopes**—were developed for viewing the upper part of the gastrointestinal tract. **Colonoscopes** are used for examining the colon and may be utilized for the early detection of carcinoma of the large bowel. Benign polyps may be detected and removed. For this purpose a special metallic snare is inserted through the instrument channel, passed over the polyp, and a high-frequency current used to heat the wire and remove the polyp (electroresection). **Bronchoscopes** are thinner, as they are used to visualize the thin bronchi. Bronchoscopes have also been used for removal of foreign bodies, and for the surgical removal of tumors. **Laparoscopes** of diameter 5–10 mm are inserted into the body through an incision in the navel. They have been used for the removal of the gallbladder.

2. Thin and Ultrathin Endoscopes

Recently there has been progress in developing endoscopes of very small diameter. Flexible endoscopes of diameter of about 0.5–2 mm may incorporate up to 10,000 fibers, each of diameter of a few micrometers. The length of the endoscope is about 1 m, and the resolving power is high enough to see a thin polypropylene suture inside a blood vessel. Some of these thin endoscopes are intended for cardiology and they are then called angioscopes. Such angioscopes have been inserted through blood vessels into the heart and used for examining heart valves. They have also been inserted into the coronary arteries and used for viewing atherosclerotic plaque. Thin, fiberoptic endoscopes have been used in fetoscopy—imaging of the fetus inside the womb during pregnancy. These endoscopes can be inserted through a small incision in the abdominal wall or through the uterine entry. They may provide important diagnosis in the early stages of pregnancy, where the resolution of ultrasound is insufficient. Ultrathin endoscopes, of diameters smaller than 1 mm, can be used for imaging of other body parts. For example, rigid endoscopes have been inserted into teeth and used for imaging of root canals.

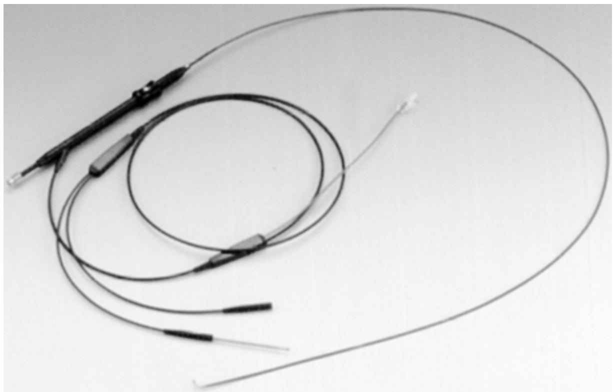


FIGURE 8 A picture of an ultrathin endoscope.

A picture of an Olympus ultrathin endoscope of diamter less than 1.0 mm is given in Fig. 8. Such endoscopes have been successfully used during balloon angioplasty procedures (PTCA, as explained in Section VII.E on cardiology). Figure 9 shows the results of standard angiography (upper three images) and of fiberoptic endoscopy (three lower images), carried out during the same procedure. The angiograms and the endoscopic images were obtained before the procedure (Fig. 9A) and after the procedure (Fig. 9B & C). The angiograms can only show the shadow of an opaque fluid inside an artery. On the other hand, the endoscopic images can show the actual plaque blocking the blood vessel, and its removal.

V. FIBEROPTIC MEDICAL DIAGNOSTICS

There is a need to improve some of the diagnostic medical techniques. At present, blood samples are sent to a remote laboratory for analysis. This laboratory may be far from the patient and the physician and there are bound to be delays or even unintentional errors in the clinical chemical results. There is a concentrated effort to use miniaturized electronic devices as sensors that could perform chemical analysis inside the body, in real time. Fiberoptics offers an alternative method for performing medical diagnostics inside the body of a patient. In principle, this method may prove to be sensitive, reliable, and cost effective.

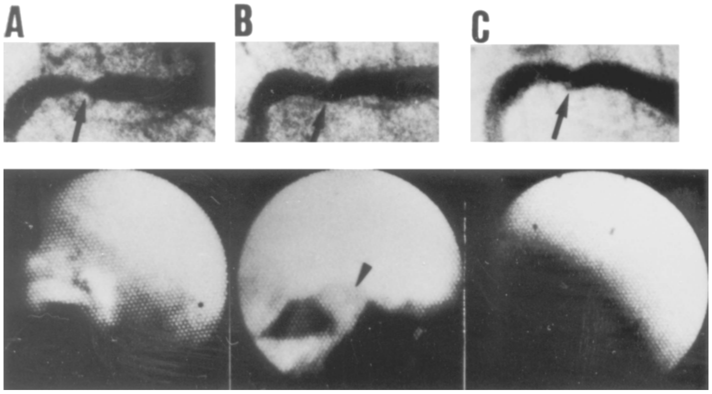


FIGURE 9 Image obtained before and after a balloon angioplasty (PTCA) procedure. (A) Top: The angiogram shows a shadow of the blocked artery. The black arrow points to the blockage. (B) Bottom: The endoscopic image shows the actual blockage. (B) & (C) Top: The angiograms show that blockage was removed and blood flow was resumed. (B) & (C) Bottom: The endoscopic images at the bottom also show that the blockage was removed.

A. Diagnostic Systems

A typical fiberoptic sensor system that could be used for medical diagnostics is shown schematically in Fig. 10. The laser beam is coupled into the proximal end of an optical fiber and is transmitted to the distal end, which is located in a sampling region. Light is transmitted back through the same fiber (or a different fiber) and is then reflected into an optical instrument for optical analysis. The fiberoptic sensors fall into two categories: direct and indirect, as shown in Fig. 11. When using direct sensors, the distal end of the fiber is bare, and it is simply inserted into the sampling region. Indirect sensors incorporate a transducer at the distal end (sometimes called an optode, in analogy to the electrical electrode). The laser light interacts with the transducer, which then interacts with the sample. Each of these categories may be further divided into two: physical sensors and chemical ones. Physical sensors respond to some physical change in the sample such as temperature or pressure, while the chemical sensors respond to changes that are chemical in nature, such as pH. We will discuss each category separately and illustrate their operation with a few examples.

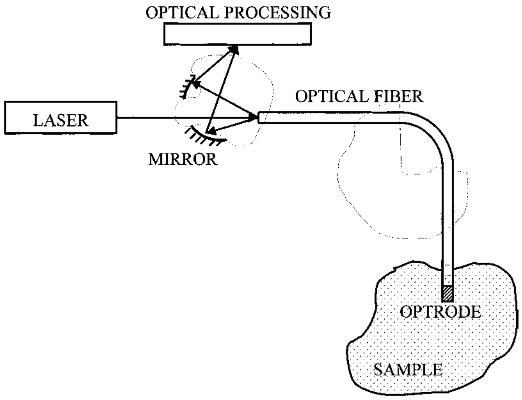


FIGURE 10 A fiberoptic sensor for medical diagnostics.

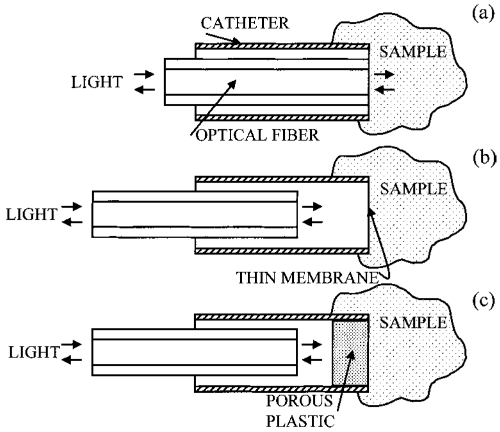


FIGURE 11 Fiberoptic sensors. (a) A direct sensor. (b) An indirect physical sensor. (c) An indirect chemical sensor.

B. Direct Sensors

1. Physical Sensors

Blood velocity may be measured by using a fiberoptic technique called laser Doppler velocimetry (LDV). A glass fiber is inserted through a catheter (plastic tube) into a blood vessel. A He-Ne laser beam is sent through the fiber, directly into the blood. The light scattered back from the flowing erythrocytes (blood cells) is collected by the same fiber and transmitted back. This scattered light is shifted in frequency (with respect to the regular frequency of the He-Ne) because of the Doppler effect. The frequency shift is proportional to the velocity. Using the LDV technique, instantaneous arterial blood velocity measurements are performed. Blood-velocity profiles at

an arterial stenosis (narrowing) are also studied with the same technique.

Blood flow may also be measured by a dye dilution technique. As an example, sodium fluorescein may be injected into the blood. This dye has a yellow-green luminescence when excited by blue light. An optical fiber is inserted into an artery, and the same fiber transmits both the excitation light and the emitted luminescence. The dye dilution in the blood can be deduced from the decrease in luminescence with time. This in turn may be used to calculate blood flow.

There are other physical effects that can be directly measured with optical fibers. For example, light scattering from tissues can be easily measured. The scattering from diseased tissues may be significantly different from that from healthy tissue.

2. Chemical Sensors

Oxygen in blood is carried by the red blood cells. Oxygen saturation is the ratio (in percentage) between the oxygen content in a blood sample and the maximum carrying capacity. In arteries the blood is normally more than 90% saturated, and in the veins only about 75%. There is a large difference between hemoglobin and oxyhemoglobin in absorption of light at $\lambda = 660$ nm. On the other hand, there is little difference at $\lambda = 805$ nm. Light at $\lambda = 660$ nm, emitted from a semiconductor source (or incandescent lamp), is sent through one optical fiber. The light scattered back from blood is transmitted through a second fiber, and its intensity I_{660} is measured. Two other fibers are used for monitoring the intensity I_{805} for calibration. The ratio I_{660}/I_{805} is used to determine the oxygen saturation in blood. Measurements have been routinely performed through a thin catheter of diameter 2 mm.

C. Indirect Sensors

1. Physical Sensors

In this case, a transducer (optode) is attached to the distal tip of an optical fiber catheter, in order to perform physical measurements. Two measurements are of special importance: temperature and pressure.

There are several schemes for measuring temperature. One scheme is based on the change in luminescence of a phosphor with temperature. The phosphor powder is attached to the distal end of a plastic clad silica fiber. A pulse of UV light is sent through the fiber and excites the phosphor. The visible luminescence is returned through the same fiber, and its decay time is measured. The decay time is dependent on temperature, and its measurement is translated into a temperature reading. The accuracy is better than 0.1°C near room temperature.

Pressure may be measured via mechanical transducers attached to the optical fiber. For example, a reflective surface may be attached to the distal end of a fiber by flexible bellows [Fig. 11(b)]. Light sent through the fiber is reflected back through the same fiber. The light output depends on the shape of the reflective surface, which in turn depends on pressure. Both temperature and pressure have been measured *in vivo* inside blood vessels.

2. Chemical Sensors

In this case, the miniature transducers, which are attached to the end of the optical fiber, are sensitive to chemical changes in the sample of interest (e.g., blood). The basic design of an indirect chemical sensor is shown in Fig. 11(c). A special reagent is trapped inside a porous polymer or sol-gel layer. Light is sent through one fiber, and reflected light or luminescence from the reagent is transmitted back. The reagent is allowed to interact with blood, for example, through diffusion. This, in turn, is manifested as a change in luminescence (or change in reflected light). Sensors for measuring pH are based on a dye indicator. Some dyes luminesce under UV excitation, and the luminescence intensity is determined by the pH of the sample. For other dyes the absorption spectra are determined by the pH. In practice, fiberoptic sensors have been used for measuring pH of blood in the physiological/pathological range (pH = 6.8–7.7) with accuracy better than 0.01 pH units.

A different sensor, of similar construction, is used for monitoring the partial pressure of oxygen pO_2 in blood. In this case, one may again use a dye that fluoresces. In some dyes the fluorescence emission decreases with increase in pO_2 (quenching). The fluorescence intensity is therefore a direct measure of pO_2. Similar sensors have also been used for measuring the partial pressure of CO_2, pCO_2, in blood.

The feasibility of using optical fibers as biomedical sensors has been established. Some technical problems (e.g., response time, calibration, and shelf life) need further attention. Some of the sensors described in Section V have already been introduced into clinical use. A new family of indirect chemical sensors incorporates biomolecules such as enzymes or antibodies. These sensors can monitor the body's levels of glucose or penicillin and may soon be used to measure metabolic substances, toxins, and microorganisms in the body.

VI. INTEGRATED FIBEROPTIC SYSTEMS

A. Power Fibers for Medical Systems

Therapeutic applications of lasers, and in particular surgical applications, call for using relatively high laser power.

Typically, an average laser power of 10–100 W is required. With the advent of low-loss fibers it is possible to transmit such power levels through thin optical fibers. Power fibers may replace the cumbersome articulating arms for delivering the beam from the laser to the operating site. These power fibers may also be inserted inside the human body through natural orifices, through catheters, or through flexible or rigid endoscopes. Laser beams sent through the fibers may then be used for a variety of medical applications: coagulation, therapeutic heating, tissue welding, ablation, incision, among others.

The three lasers that have been commonly used in the past for laser surgery are the CO_2, the Nd:YAG, and the Ar ion lasers. These are also the lasers that were tried first with fiberoptic delivery systems. The Ar laser radiation ($\lambda = 514$ nm) is in the visible spectral range, and its radiation is readily transmitted by pure silica glass (i.e., quartz) fibers. The same is true for the Nd:YAG laser, whose radiation is in the NIR ($\lambda = 1.06\,\mu$m). Fibers of diameters 0.1–0.2 mm and of lengths 1–2 m have often been used. The power levels continuously transmitted through the fibers have been up to 10 W for the Ar ion laser, and up to 60 W for the Nd:YAG laser. The CO_2 1aser radiation is in the mid-IR ($\lambda = 10.6\,\mu$m), and it can be transmitted by hollow fibers or by polycrystalline halide fibers (e.g., ones made of silver halide). In both cases, fibers of diameter of about 1 mm and length of about 1 m have been used. Power levels of tens of watts have been continuously transmitted through these fibers.

As for other lasers that are used now for medical applications, the GaAs type laser radiation or the Ho: YAG laser radiation can also be sent through silica glass fibers. The transmission of the excimer laser radiation is more difficult, first because of the short wavelength and second because of the need to transmit very short pulses of high peak power. Specially manufactured quartz fibers are normally used. For the transmission of Er:YAG laser radiation, one can use fibers made of sapphire.

There are still many hurdles. Some of the problems involved in using optical fibers for power transmission are as follows.

1. *Damage to the ends of the fibers*. The power density (W/cm^2) at each end of the fiber is very high. It may easily reach values of 10^4 W/cm^2. Any defect on the fiber end may increase the absorption of laser beam on the end face. With these high power densities, the end face will be damaged.
2. *Bending*. With most fibers today, one cannot bend the fibers beyond a certain value. Bending increases the loss, and again, the fiber may be damaged at the bend.
3. *Divergence*. The light emitted from fibers is not collimated, as is a regular laser beam. It is highly divergent, and the NA determines the divergence. When transmitting the beam through the fiber, the distal end of the fiber has to be kept clean, and it must not touch the tissue. If this end is held 2–3 mm away from the tissue, the power density at the tissue may be too low for incision.
4. *Dry field*. Blood will coagulate under laser radiation. In order to prevent coagulation during fiberoptic laser surgery, one has to replace the blood near the distal end of the fiber by saline solution or by transparent blood substitute, or push the blood back using pressurized CO_2 gas.

The power handling capabilities of optical fibers has increased as a result of improvements in the properties of the fibers and better handling of the fiber end faces. Many of the problems associated with coupling of high-power laser beams into fibers have been solved. As a result, special fibers have been added to medical lasers as standard items for power transmission.

B. Laser Catheters

A catheter is a flexible plastic tube that can be inserted into various areas in the body. An optical fiber can be easily threaded into such a catheter, and laser energy delivered through the fiber can be used for diagnosis, therapy, or surgery. Many of the new devices are multichannel catheters in which an optical "power" fiber is inserted in one of the channels. Cooling liquid may also be injected into this channel. Laser power sent through such a fiber could be used either for tissue removal (i.e., surgery) or for therapeutic heating, for instance, laser-induced laser therapy (LITT). Other channels are used for a diagnosis fiber, for irrigation of tissue with saline solution, or for introducing drugs. A metal guide wire is often inserted in one of the channels to facilitate easy insertion of the catheter into the body. Recent progress in such catheters has led to their surgical and therapeutic use in cardiology, gynecology, orthopedics, and other medical disciplines. Several examples are discussed below.

C. Laser Endoscopes

With the development of new flexible fiberscopes of very high optical quality and the concurrent development of optical fibers for laser power transmission, the road was clear for developing novel endoscopes. The endoscope would consist of several channels. One channel would include a complete fiberscope, consisting of an imaging bundle and illumination light guides. Another channel would

include optical fibers for diagnostic purposes, and the distal end of these fibers might even include transducers (optodes). A third channel would include a power fiber, for laser power delivery. A fourth channel could be used for injecting liquids such as drugs, dyes, or saline solution. This (or another) channel could be used for pumping out liquids or for the aspiration of gases.

Some of these systems incorporate novel devices such as miniature ultrasound devices that provide three-dimensional (3D) imaging. This is particularly useful in cardiology for 3D imaging inside blood vessels. Laser endoscopes have reached the stage where they are being used clinically in several medical disciplines such as otolaryngology, gynecology, gastroenterology, and brain surgery. Some examples are mentioned in Section VII.

D. Robotic Systems

Incorporating complex computer controls, which make it a robotic system, could enhance the effectiveness and safety of surgical endoscopic systems. In a simple version of the system, the thin articulated arms of a robot are equipped with surgical tools. These are inserted into the body through a small incision. The physician does not directly operate the surgical instruments. Instead, the physician operates joysticks at the robot's computer control, which is located near the patient. These joysticks, in turn, control the placing and the operation of the surgical instruments using the sophisticated electronics and software of the system. A thin endoscope is also inserted into the body. This endoscope could be a flexible or rigid fiberoptic endoscope, with two imaging bundles, or it could incorporate two tiny CCDs. In both cases, it would be possible to generate a 3D image of the treated site, which is important for the control of the surgical tools inside the body. The accuracy of the operation of such a robotic system can, in principle, be very high. In the future, the articulated arms of the robot could be replaced by optical fibers, which will provide both diagnosis and therapy (or surgery), and the whole system will be endoscopic.

VII. LASER–FIBEROPTIC SYSTEMS AND THEIR CLINICAL APPLICATIONS

An increasing number of physicians are using integrated laser–fiberoptic systems, such as laser catheters or laser endoscopes, in a clinical setting. A few examples illustrate the use of laser catheters and endoscopes for clinical applications. The sections on cardiology and on cancer diagnosis and therapy illustrate the enormous potential of laser–fiberoptic techniques.

A. Gastroenterology

One of the first uses of the fiberoptic laser endoscope was in treating bleeding ulcers. The laser light could photocoagulate blood and therefore cause hemostasis (cessation of bleeding). Among the three important lasers (CO_2, Nd:YAG, and Ar), some physicians prefer the Nd:YAG laser because it penetrates deep into tissue and its effects are not localized at the surface. Others have used the Ar ion laser for controlling gastric hemorrhage (bleeding). In both cases, the laser beam is sent through thin silica "power" fibers. These procedures have been performed on patients with bleeding ulcers in the stomach, the esophagus, or the colon, and the results have been very good.

B. Urology

An Nd:YAG laser beam, sent through the power fiber of a laser endoscope, is used for destroying small tumors in the urinary bladder. In other cases, larger tumors are electroresected (using the endoscopic snare described earlier), and Nd:YAG radiation is then used for coagulation. Again, the deep penetration of Nd:YAG in tissue may be advantageous.

Laser catheters based on dye lasers and Nd:YAG lasers have been used for laser lithotripsy, the shattering of stones inside the body. In this case, the catheter is inserted into the body and the distal tip of the power fiber is brought in contact with the stone. Dye laser, Ho:YAG, or frequency-doubled Nd:YAG laser pulses sent through these quartz fibers are absorbed in the stone surface and generate hot plasma. The resulting shock waves rapidly shatter the stone. This procedure is particularly useful in cases where other methods such as shock wave lithotripsy cannot be applied. Laser lithotripsy has been successfully applied on patients for the removal of urinary stones or biliary stones. (Comment: an identical procedure has been used in otolaryngology for salivary gland calculi.)

With aging, the prostate of the human male enlarges and presses on the bladder or the urethra. This condition, called benign prostatic hyperplasia, results in poor urine flow and large residual urine volume. The prostate can be resected surgically using a cauterizing tool inserted through the urethra. This transurethral resection of the prostate (TURP) is a very common surgical procedure, but has many undesirable side effects. Alternatively, a laser catheter could be inserted through the urethra, and placed inside the prostate. A visible laser beam sent through the power fiber can then be used to ablate the prostate, in a procedure called visual laser prostatectomy (VLAP). Alternatively, laser-induced laser therapy can be used for the treatment of the prostate gland. A beam of Nd:YAG laser

or diode laser could penetrate deep into the gland, causing coagulation and shrinkage. These minimally invasive procedures are still under investigation, in an attempt to better control the thermal effects.

C. Gynecology

A hysteroscope is inserted into the uterus and carbon dioxide gas or a transparent fluid is injected through the endoscope to expand the cavity. This allows the physician to clearly see the internal structure. The surgeon can then use mechanical tools or a laser beam sent through a power fiber to treat polyps, scar tissue, or other abnormalities.

Laparoscopes are often used to view the uterus, fallopian tubes, and ovaries. Abnormalities, such as tube blockage or adhesions, can be corrected, again using mechanical tools or fiberoptic laser surgery. Alternatively, thin fiberoptic endoscopes can be inserted into the fallopian tubes, through the cervix and the uterus. These can be used to view directly tube blockages that may lead to infertility. Fiberoptic laser surgery could be used in the future to remove the blockages.

D. Fetal Diagnosis and Therapy

Small diameter fiberoptic fetoscopes can be inserted into the bodies of pregnant women for direct visualization of the fetus. They are useful in cases of complications, where ultrasound imaging cannot provide sufficient information. Power fibers inserted through these endoscopes can be used for laser surgery on the fetus. For example, in 10–15% of identical twins one fetus is passing blood to the other fetus over abnormal blood vessels in the placenta. This condition is called twin-to-twin Transfusion syndrome (TTTS) and in this case, the mortality rate of the fetuses is very high. Nd:YAG laser radiation sent through a power fiber is used to destroy the connecting blood vessels, under direct endoscopic imaging. The success of these preliminary experiments will pave the road for other applications of surgical fetoscopy.

E. Cardiovascular and Heart Surgery

In the cardiovascular system a major role is played by the coronary arteries, which supply blood to the heart muscle, and a myocardial infarction may result if a coronary artery becomes occluded. A common problem with the arteries is a build up of atherosclerotic plaque on the interior walls. The plaque, which contains fatty material and calcium, starts blocking the coronary arteries, and the blood flow through them is reduced. This results in angina pectoris, a condition that afflicts millions of people.

1. Laser Angioplasty Using Laser Catheters

A common technique for cardiovascular diagnosis is an X-ray study of the arteries called angiography. A thin catheter is inserted through an incision in the groin (or the arm) and is pushed through the arterial system until it reaches the coronary arteries. Then an X-ray-opaque liquid is injected through the catheter, and the shadow of this liquid is examined by X-ray imaging. A blockage in the blood flow can thus be examined. If the coronary artery is only partially blocked, the situation can sometimes be improved by using a method called balloon angioplasty. The more exact name for this method is percutaneous transluminal coronary angioplasty (PTCA). This method makes use of a special catheter that includes a tiny balloon at its end. The tip of the catheter is inserted through the partially constricted artery, the balloon is inflated under high pressure (10 atm), and the blockage is reduced. Unfortunately, PTCA can be successfully used on only a limited number of patients. In many cases, if the arteries are blocked, one has to resort to a surgical procedure called coronary artery bypass grafting (CABG). A vein is removed from the vicinity of the blocked artery or from some other part of the body (e.g., the leg) and is implanted in parallel to the blocked artery. This operation requires an open-heart procedure, which is traumatic, risky, and expensive.

With the development of optical fibers that could deliver high laser energy, during cardiac catheterization, when substantial blockage of the coronary arteries is observed, an optical fiber could be inserted through the catheter. A laser beam sent through the fiber could then be used to vaporize the plaque and open a clear channel through which blood flow can then resume. This procedure is called laser angioplasty or vascular recanalization, and a schematic cardiovascular laser catheter and its operation is shown in Fig. 12.

Laser angioplasty of the coronary arteries has been successfully accomplished with laser catheters based on the excimer laser and on pure silica power fibers. Excimer lasers, emitting in the UV ($\lambda < 200$ nm), were selected, because they ablate tissue without causing thermal damage. The tip of the catheter is guided using X-ray fluoroscopy, and the excimer laser is used to vaporize plaque with little thermal damage. Thousands of patients have been treated by this method. Laser catheters based on other lasers and fibers have also been tried. One of the problems not yet fully solved is that of monitoring and control of the laser beam during the procedure. Some of the laser catheters use fiberoptic sensor techniques to distinguish between plaque and normal arterial wall and to prevent perforation of the arteries. Progress in the use of laser catheters for laser angioplasty of the coronary arteries has been slower than hoped, due to the complexity of the procedure.

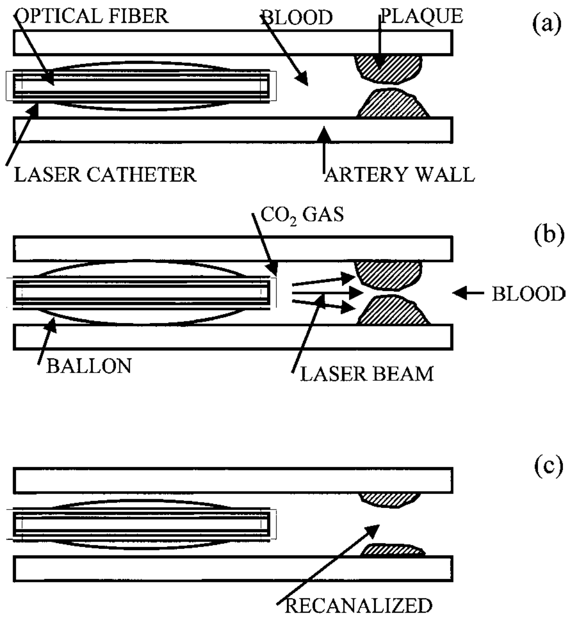


FIGURE 12 Laser catheter for recanalization of arteries. (a) A laser catheter inside an artery blocked by plaque. (b) A balloon is inflated, and CO_2 gas (of saline solution) pushes the blood away. Laser beam vaporizes the plaque. (c) Artery recanalized, and blood flow resumed.

2. Transmyocardial Revascularization

Restriction of blood flow to the myocardium causes angina pectoris, and also puts the patient in danger of a heart attack. Blood supply for the heart muscle in reptiles is not based on coronary arteries. Instead, there is a perfusion of blood through the heart walls into the muscles. Physicians have tried for many years to imitate this situation in the human heart. The idea was to drill holes through the myocardium, so that oxygenated blood from the heart chambers will seep into the heart muscle. The procedure is called transmyocardial revascularization (TMR). Drilling holes using mechanical tools did not succeed. In early experiments in the 1970s high-energy CO_2 laser pulses were tried. The goal was to try to help patients who could not undergo heart bypass surgery or angioplasty. Extensive studies began in the 1990s, using both CO_2 lasers and then Ho:YAG lasers. In these experiments the chest cavity was opened to allow access to the heart, and high-energy laser pulses were used to cut holes through the outer side of the myocardium wall. The results of the experiments are still controversial, but there have been reports that the chest pains were relieved in many patients.

The "direct" TMR procedure mentioned above is an invasive procedure, involving open-heart surgery. Instead, a minimally invasive procedure, the fiberoptic TMR, can be used. In this procedure a laser catheter is inserted through an artery in the leg and guided into the left ventricle. The distal tip of the fiber is brought in contact with the inner side of the myocardium. Short laser pulses are used to drill channels, roughly halfway through the myocardium wall. Clinical investigations, using various lasers, are in progress in several medical centers.

3. Closed Chest Endoscopic Surgery on a Beating Heart

Cardiac surgeons would like to replace the open-heart CABG surgery with a closed chest, minimally invasive, bypass operation. Recent advancements have been achieved with the help of the robotic surgery, using a system similar to the one described in Section VI.D. This computer controlled robotic system provides the exquisite accuracy needed for the bypass coronary surgery. The articulated arms of the robot, equipped with surgical tools, are used to perform the whole endoscopic CABG operation on a beating heart, without arresting the operation of the heart and without placing the patient on a heart–lung machine. The same robotic system has been used for the replacement of the mitral valve in patients. Early clinical experiments have demonstrated the feasibility of using the robotic assisted system in cardiology and heart surgery.

F. Cancer Diagnoses and Photochemotherapy

It has been known since the 1940s that certain compounds, called porphyrins, are preferentially concentrated in malignant tumors, with respect to healthy tissue. Porphyrin fluoresces under UV excitation, and therefore by illuminating tissue, one may distinguish malignant tumors from benign tissue. In the I960s, a compound called hematoporphyrin derivative (HPD) was found to have even better properties than porphyrin. HPD and other compounds are now used for cancer diagnosis and therapy. This method is based on three interesting properties of these compounds:

1. *Selective retention*: In practice, HPD may be injected into a patient, and after a few days this dye concentrates only in cancer tissue.
2. *Diagnosis*: If a tissue area is now illuminated with a suitable UV source, malignant tissue will emit a characteristic red light. The fluorescence of HPD is mostly in the red part of the spectrum, with two prominent emission peaks at 630 and 690 nm. This fluorescence can be excited by UV or by blue light, but excitation at around 400 nm gives rise to the highest fluorescence. The fluorescence efficiency (emitted red power divided by excitation UV power) is fairly low, and therefore lasers are required for

excitation. A krypton laser emitting about 0.25 W at 413 nm is suitable for this application. In order to see this red emission one has to attach to an imaging system an optical filter that transmits at 630 nm and blocks the exciting light at 413 nm.

3. *Photodynamic therapy*: It was found that the use of HPD with lasers could serve not only for diagnosis of cancer but also for therapeutic purposes. If rather than illuminating a tumor with UV light, one uses red light ($\lambda = 630$ nm) of sufficient energy, the results are strikingly different. Red light penetrates deeply into the tissue. HPD absorbs this red light, a series of photochemical reactions occur, and the end result is the release of some photoproduct (probably singlet oxygen), which kills the host malignant tissue. The method itself is called photodynamic therapy (PDT). This is actually photochemotherapy, because it involves the administration of a drug and the use of optical radiation for triggering the process that cures the disease. Roughly 10–50 mW/cm^2 of red light is needed for this photochemotherapy.

The diagnostic method mentioned in (1) can be readily adapted in endoscopy. A special endoscope could incorporate a quartz fiber for UV illumination and a red transmitting filter in front of the imaging fiberscope. With such an endoscope, tumors inside the body could be irradiated by krypton laser light (or even by a Xe lamp with a UV transmitting filter), and using the red image formed, one could locate malignant tumors. This may be used for early detection of lung cancer, when the malignant tumors are too small to be detected by chest X-ray or by computed tomography.

Photodynamic therapy, mentioned in (2), is also adaptable for use in conjunction with fiberoptic systems. A high-intensity red light can be transmitted through a quartz fiber, and delivered directly inside the tumor. This light may then selectively destroy cancer cells. In the past, a Nd:YAG laser pumped or Ar laser pumped dye laser, emitting at $\lambda = 630$ nm, was used for this purpose. There has also been progress with the development of miniature semiconductor lasers, which emit continuously several watts at 630 nm.

The endoscopic diagnosis and treatment of cancer has been reported in the esophagus, lungs, larynx, prostate, and cervix. Other photosensitive drugs are being tested, including benzoporphyrin derivative and phthalocyanines.

Laser endoscopic photochemotherapy has been proposed for noncancer applications, such as the treatment of age-related degenerate macula. Another example is photodynamic laser angioplasty, which would be used to treat coronary artery disease. Other examples are the treatment of rheumatoid arthritis, in orthopedics, and the treatment of the endometrium, in gynecology. Different photosensitive drugs, such as aminolevulinic acid (ALA) and lutetium texaphyrin, have been also tried for these applications.

VIII. NOVEL FIBEROPTIC MEDICAL SYSTEMS

During the last 5 years there has been rapid advancement in the field of biomedical optics. The basic and preclinical research was described in Sections I–VII. There is a plethora of novel methods and medical systems and many of these are now being tried clinically. In this section we discuss a few of these new developments.

A. Time Resolved Phenomena

There is great interest in the early detection of breast cancer. The problem is to detect tumors of a size of a few millimeters in tissue of total thickness 40–100 mm, and to try to distinguish between malignant and benign tumors. The traditional methods such as X-ray mammography or ultrasound imaging do not always provide the necessary resolution, and they cannot distinguish between the types of tumors. Magnetic resonance imaging is a very powerful method but it is cumbersome and expensive. If there is a small cancerous tumor inside the breast, the optical absorption in it is higher, due to higher density of blood vessels, and the optical scattering is also higher, due to changes in the refractive index. Optical techniques could provide a way of distinguishing between two tissues. Such techniques would be useful, not only for breast cancer but for many other applications. One of these optical methods is schematically described here.

Let us consider an NIR laser that emits very short pulses (of the order of a few femtoseconds) in a narrow beam. The beam impinges on one side of the breast and it encounters scattering and absorption. Photons that are transmitted are detected on the other side of the breast by a fast optical detector. There are three types of photons: (i) ballistic photons which pass through the tissue in straight lines, without scattering, (ii) “snake” photons which pass through the tissue in a zigzag path, with little scattering, and (iii) diffuse photons, which suffer multiple scattering. The ballistic photons will arrive at the detector much faster than the other photons. One may therefore place a very fast shutter in front of the detector; use the shutter to let only the ballistic photons through, and then turn the shutter off (i.e., optical gating). If the narrow beam is scanned across the breast, there will be a noticeable difference if it is transmitted through a tumor (which has higher absorption and scattering coefficients) or through healthy tissue. This

transillumination technique may therefore give rise to an image of a tumor. This time resolved imaging technique is somewhat similar to X-ray imaging, but it may provide higher resolution and it does not suffer from the health hazards of X-rays.

Instead of scanning the beam across the breast one may use two bundles of optical fibers. In each of them the individual fibers at the proximal end are tightly arranged. At the distal ends the fibers are spread, and they surround the breast. A very short pulse emitted from a diode laser or a dye laser is focused on one of the fibers in the proximal end of one of the bundles. The optical energy transmitted through the breast is collected by the distal ends of the second fiber bundle and transmitted to the proximal end. The energy collected at each of these fibers is measured, one by one, using a fast optical detector. The focused laser beam is moved to a different fiber in the first bundle, and the process is repeated. From the analysis of the results one may obtain the image of a tumor in the breast.

This and similar imaging techniques have been tried for imaging of breast cancer and brain cancer. Early results of this technique are very promising. There are many technical hurdles that have to be overcome before this method could be used routinely. It is expected that the spatial resolution will be of the order of 1 mm, adequate for the early detection of breast cancer.

B. Optical Coherent Tomography

Ultrasound imaging is based on bouncing ultrasound waves off tissue and determining the delay time of waves reflected from internal structures in the tissue. This can be translated to 2D images with a resolution of a few hundred microns. Optical coherence tomography (OCT) is based on the bouncing of light waves from tissue, and measuring how long it takes for the backscattered light from microstructures in tissues to return. Because of the very short delay times involved, conventional electronics cannot be used, and one has to apply other techniques. In its simplest form, an OCT system is similar to a Michaelson interferometer, which includes an NIR laser source with a relatively broad band of wavelengths. One beam from the interferometer focuses on a tissue sample, whereas a second beam provides a reference. The reflected light from some depth in the sample interferes with the reference signal, only if the depth is identical to the length of the reference arm. The reference arm is scanned and a detector records the interference patterns between the two beams, which are actually generated by reflections from different depths in the tissue. The analysis of the interference patterns of the two beams can be used to obtain 2D images of the tissue, at a depth of a few millimeters, with a resolution of a few micrometers.

The first application of OCT was in ophthalmology, where fine structures in the eye (i.e., macular hole in the retina) have been observed. The system was then adapted for endoscopic applications. The interferometer was equipped with optical fibers, and one of them was inserted into a thin endoscope. By attaching a rotating prism to the distal tip of the fiber, or by scanning the fiber, one may obtain a cross sectional image. Endoscopic OCT has been tried clinically to study the gastrointestinal tract and the esophagus. It is being tried as well for intravascular imaging, in cardiology, to study plaque inside the arteries. Endoscopic OCT offers physicians the ability to view microscopic structures inside the body in real time (even at a video rate) and with very high resolution. This is a form of "optical biopsy" which provides images of tissue *in situ*, with no need to remove tissue specimens for examination.

IX. OUTLOOK

During the past decade there has been rapid progress in the development of optical fibers and fiberoptic systems. Foremost was the development of silica-based glass fibers, which have extremely low transmission losses. These fibers are the major building blocks of future communication systems. Other optical fibers were developed for special applications, such as fibers with good transmission in the mid-IR or in the deep-UV parts of the spectrum. Among these we may also include power fibers that are capable of transmitting laser beams of high intensities. With the development of the various optical fibers, there has been an increasing recognition of the enormous possibilities for utilizing them in medical applications. Three major uses of fiber—optical endoscopy, diagnostics, and power transmission—have been reviewed here.

In the past, endoscopes were rigid, and their uses were limited. A drastic change occurred when fiberscopes (especially the flexible ones) were included in endoscopes. Rigid endoscopes are still widely used in many applications because they do provide an excellent image, it is easier to steam sterilize them, and they are cheaper. As mentioned earlier, there are videoscopes based on CCDs that are also used in endoscopy. These are still bulky and expensive, but progress is being made to reduce their size and cost.

An endoscopist could get a clear image of an internal organ, and perform surgical operations such as biopsy or the removal of the gallbladder, by inserting medical instruments through ancillary channels. A computer, such as the one used in robot–assisted surgery, could control the surgical instruments. The first operations carried out with a laser-plus-robot system were the minimally invasive coronary bypass operations (with rigid telescope).

Robot–assisted endoscopic operations have also been used in gynecology, for the reconnection of fallopian tubes, and in general surgery, for the laparoscopic removal of gall-bladders (rigid).

Thin fiberscopes of diameter 2–3 mm have been inserted into blood vessels and used to visualize atherosclerotic plaque within the coronary arteries, or to view the inner parts of a fetus inside the womb. There has been further progress in the development of ultrathin flexible fiberscopes whose outer diameter is less than 1 mm. With further development, such fiberoptic endoscopes will undoubtedly have a growing role in medicine.

Optical fibers can be used for diagnosis by inserting them into the body and making physical and chemical measurements through them. As direct sensors, the fibers serve simply as light guides for transmitting laser light into the body and back. Such direct sensors have been used for measuring blood flow or for the early detection of cancer. In a different mode of operation, tiny transducers (optodes) are attached to the fibers to form indirect sensors. With these optodes, physical measurements, such as blood pressure and temperature, or chemical measurements, such as pH or pO_2, can be carried out. Currently, blood samples are extracted from a patient and sent to the laboratory for chemical analysis. Fiberoptic techniques may bring the laboratory closer to the patient. They may also enable fast and repetitive chemical analysis to be performed at the patient's bedside or during an operation. The use of optical fibers as sensors may simplify some of the diagnostic techniques and make them more reliable and less costly.

Recently there has also been progress in the development of power fibers, which could transmit relatively high laser power. Pure silica fibers have been used for the transmission of Nd:YAG, GaAs, Ar-ion, and excimer laser beams, and hollow or crystalline IR fibers for the delivery of CO_2 laser beam. Power fibers would undoubtedly replace articulating arms in surgical laser systems. Such fibers, inside laser catheters, may be inserted into the body and used to perform surgical operations, without necessitating a large incision. The enormous advantages of using fiberoptic delivery systems for surgical operations inside the body are obvious. Inserting the fibers inside the body is a minimally invasive surgical procedure and the need for a major surgical operation may be eliminated in many cases.

Finally, one may consider a compound laser endoscope (Fig. 13) that would contain several channels. One would be a fiberscope that enables the physician to see what he/she is doing. A second channel would be a fiberoptic sensor, for diagnostic purposes. A third channel would be occupied by a power fiber for transmitting high-intensity laser beams. Other channels would be used for injection

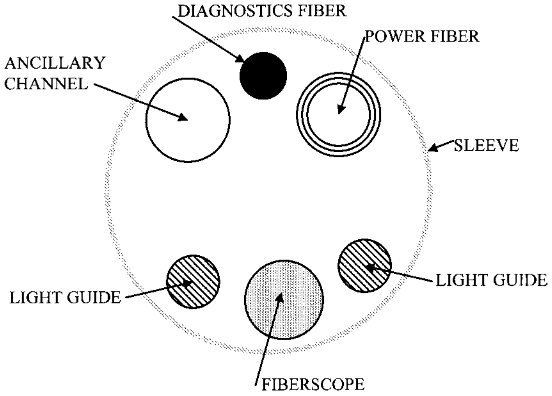


FIGURE 13 A compound endoscope, which includes a fiberscope, a fiberoptic diagnostic system, a power fiber, and an ancillary channel for irrigation or suction.

of liquids or drugs, for inserting pressurized gases, or for sucking out debris. Similar endoscopes may be used in a multitude of applications. In the future, laser endoscope systems may be inexpensive and simple to use, and they may even be used in small clinics.

No doubt fiberoptic techniques will replace more traditional procedures in the coming years.

SEE ALSO THE FOLLOWING ARTICLES

Glass • Image-Guided Surgery • Laser-Materials Interactions • Lasers, Optical Fiber • Light Sources • Microoptics • Optical Fiber Communications • Optical Fibers, Fabrication and Applications

BIBLIOGRAPHY

Carruth, J. A. S., and McKenzie, A. L. (1984). "Medical Lasers—Science and Clinical Practice," Adam Hilger, Bristol and Boston.

Katzir, A., ed. (2000). "Proceedings of the Biomedical Optics Symposia 1984–2000."

Katzir, A. (1997). "Lasers and Optical Fibers in Medicine (Physical Techniques in Biology and Medicine)," Academic Press, San Diego.

Katzir, A., ed. (1990). "Selected Papers on Optical Fibers in Medicine," Spie Milestone Series, Vol. MS 11, SPIE Press.

Niemz, M. H. (1996). "Laser–Tissue Interactions: Fundamentals and Applications," Springer Verlag.

Puliafito, C. A., ed. (1996). "Laser Surgery and Medicine: Principles and Practice," John Wiley & Sons, New York.

Sivak, M. V. (1987). "Gastroenterologic Endoscopy," Saunders, Philadelphia.

Welch, A. J., and Van Gemert, J. C., ed. (1995). "Optical-Thermal Response of Laser-Irradiated Tissue," Plenum, New York.

Wolf, H. F., ed. (1984). "Handbook of Fiber Optics: Theory and Applications," Granada Publishing, Great Britain.

Pharmaceuticals, Controlled Release of

Giancarlo Santus
Recordati Industria Chimica e Farmaceutica S.p.A.

Richard W. Baker
Membrane Technology and Research, Inc.

I. Introduction/History
II. Methods of Achieving Controlled Release
III. Important Controlled Release Products
IV. Future Directions

GLOSSARY

Biodegradable polymers Materials that undergo slow chemical degradation in the body, finally degrading to low-molecular-weight fragments that dissolve or are metabolized.

Controlled release systems Devices that meter delivery of drugs or other active agents to the body at a rate predetermined by the system design and are largely unaffected by the surrounding biological environment.

Enteric coatings pH-Sensitive materials used to delay dissolution of tablets in the acid environment of the stomach but allow rapid dissolution when the tablet reaches the more neutral pH environment of the gastrointestinal tract.

Microcapsules Small, drug-containing particles with diameters between 50 and 1000 μm.

Targeted drug delivery Delivery of a drug directly to the body site where the drug has its biological effect.

Transdermal patches Drug-containing laminates attached to the skin with an adhesive. Drug contained in the laminate migrates from the patch through the skin and is absorbed into the blood circulatory system.

Zero order delivery (of drug) Constant drug delivery over a certain period of time.

THE OBJECTIVE OF CONTROLLED drug delivery devices is to deliver a drug to the body at a rate predetermined by the design of the device and independent of the changing environment of the body. In conventional medications, only the total mass of drug delivered to a patient is controlled. In controlled drug delivery medications, both the mass of drug and the rate at which the drug is delivered is controlled. This additional level of control enhances the safety and efficacy of many drugs. Often a membrane is used to moderate the rate of delivery of drug. For example, in some devices, a membrane controls permeation of the drug from a reservoir to achieve the drug delivery rate required. Other devices use the osmotic pressure produced by diffusion of water across a membrane to power miniature pumps. In yet other devices, the drug is impregnated into a polymer material, which then slowly dissolves or

degrades in the body. Drug delivery is then controlled by a combination of diffusion and biodegradation.

I. INTRODUCTION/HISTORY

The pharmaceutical industry has produced long-acting oral medications since the 1950s. Enteric coated tablets were the first long-acting medication to be widely used. In 1952, Smith Kline French (SKF) introduced Spansules, or "tiny time pills," an over-the-counter cold medication consisting of millimeter-sized pills containing the drug and covered with a coating designed to delay its dissolution. By varying the thickness of the coating, different drug release profiles were achieved. These early products are best called *sustained release* systems, meaning that the release of the drug, although slower and more controlled than for a simple, fast-dissolving tablet, was still substantially affected by the external environment into which it was released. In contrast, the release of drug from *controlled release* systems is controlled by the design of the system and is largely independent of external environmental factors.

The founding of Alza Corporation by Alex Zaffaroni in the 1960s gave the development of controlled release technology a decisive thrust. Alza was dedicated to developing novel controlled release drug delivery systems. The products developed by Alza during the subsequent 25 years stimulated the entire pharmaceutical industry. The first pharmaceutical product in which the drug registration document specified both the total amount of drug in the device and the delivery rate was an Alza product, the Ocusert, launched in 1974. This device was designed to delivery the drug pilocarpine to control the eye disease glaucoma. The device consisted of a thin, elliptical, three-layer laminate with the drug sandwiched between two rate-controlling polymer membranes through which the drug slowly diffused. It was placed in the cul de sac of the eye, where it delivered the drug at a constant rate for 7 days, after which it was removed and replaced. The Ocusert was a technical tour de force, although only a limited marketing success. Alza later developed a number of more widely used products, including multilayer transdermal patches designed to deliver drugs through the skin. The drugs included scopolamine (for motion sickness), nitroglycerin (for angina), estradiol (for hormone replacement), and nicotine (for control of smoking addiction). Many others have followed Alza's success, and more than 20 transdermal products, delivering a variety of drugs, are now available. Alza also developed the first widely used osmotic controlled release drug delivery systems under the trade name Oros. The first billion-dollar controlled release product was an osmotic product, Procardia XL, delivering the drug nifedipine, a widely prescribed antihypertensive. Other important products launched in the last 15 years include Prilosec, a diffusion-controlled microparticle oral formulation system for the drug omeprazole, and Lupron (leuprolide) and Zoladex (goserelin), polypeptide drugs delivered from biodegradable intramuscular and subcutaneous implants. Since the first controlled release product was launched, the controlled release industry has grown to a $10–20 billion industry with more than 30 major controlled release products registered with the U.S. Food and Drug Administration. A time-line showing some of the important milestones in the growth of controlled release technology is shown in Fig. 1. Development of the technology has been reviewed in a number of monographs.

Controlled slow release of drugs to the body offers several important benefits to the patient.

- More constant drug blood levels. In controlled release products, drug delivery to the body is metered slowly over a long period, avoiding the problems of overdosing and underdosing associated with conventional medications. These problems are illustrated in Fig. 2, which shows the blood levels achieved when a relatively rapidly metabolized drug is given as a series of conventional fast-dissolving tablets. Immediately after the drug is ingested, blood levels begin to rise, reaching a peak value, after which the concentration falls as drug is metabolized. In Fig. 2 two important concentration levels are shown: the minimum effective concentration, below which the drug is ineffective and the toxic concentration, above which undesirable side effects occur. Maintaining the concentration of the drug between these two levels is critical for safety and effectiveness. Controlled release systems meter delivery of the drug to the body at a rate equal to the rate of drug removal by metabolization, so that a prolonged constant blood level is maintained. Because controlled release products use the drug more efficiently, the total amount required to produce the desired effect is often very much reduced, frequently by a factor of 10 or more.
- Improved patient compliance. A second benefit of controlled release formulations is improved patient compliance. Many studies have shown that few patients take their medications in complete accordance to physician instructions, and that the degree of noncompliance increases significantly as the complexity of the instructions increases. In one study of patients given once-per-day estrogen/progesterone contraceptive tablets in a controlled fashion, the pregnancy rate was less than 1 per 1000 patient-years. The same tablets prescribed to a normal population of women resulted in a pregnancy rate of 50 per 1000 patient-years due to noncompliance. Delivery of the similar contraceptive steroid levonorgestrel from a sustained release implant resulted in a pregnancy rate of

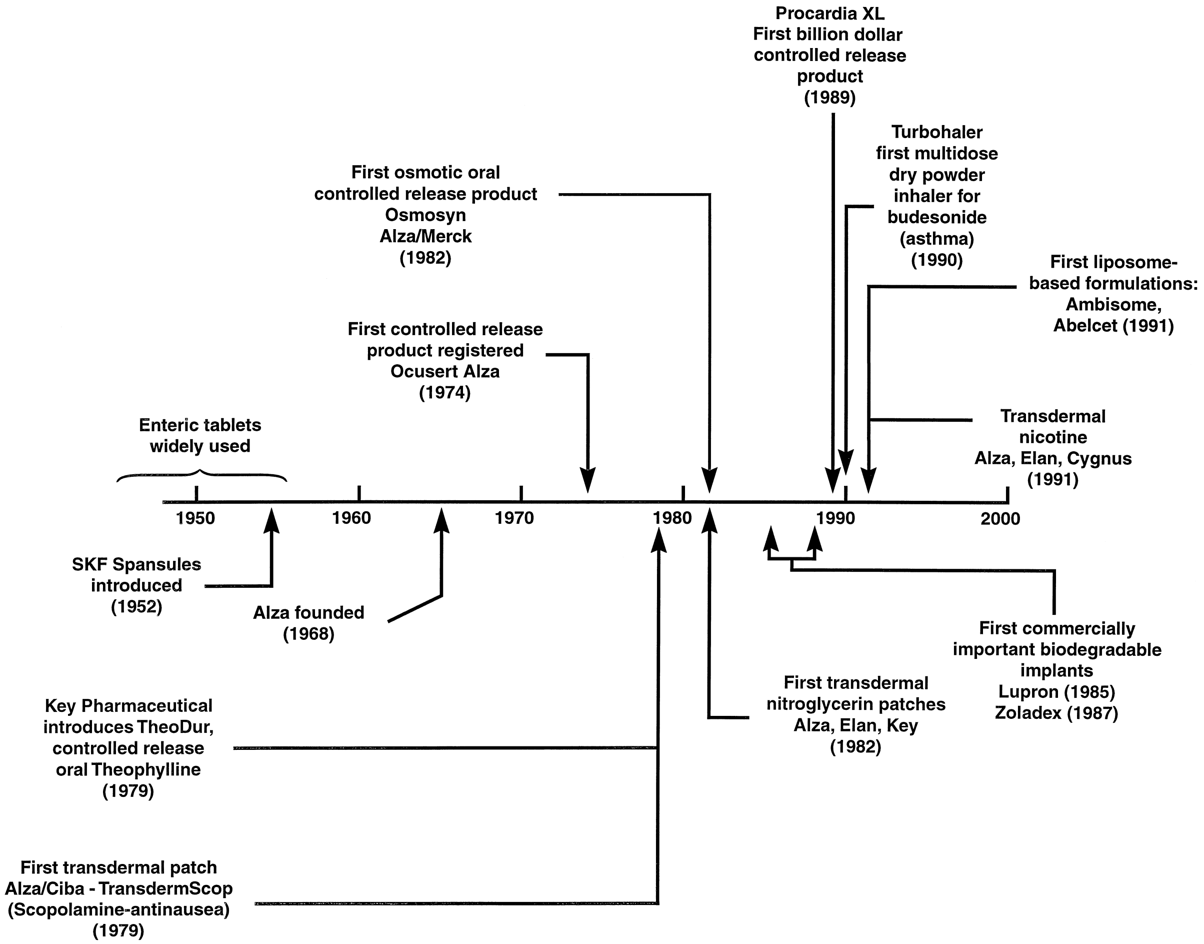


FIGURE 1 Milestones in the development of controlled release pharmaceuticals.

0.5 per 1000 patient-years. This product has the lowest pregnancy rate, that is, the highest degree of contraceptive efficiency, of any steroidal contraceptive because it eliminates poor patient compliance, the principal reason for drug failure.

• Targeted site of action delivery. A final benefit of controlled release formulation is the ability to deliver drug directly to the systemic circulation. Drugs taken orally are absorbed into the blood stream in the gastrointestinal (GI) tract. The drug then enters the portal circulation, so is first taken to the liver before entering the systemic circulation and being transported to the site of drug action. The liver's function is to deactivate toxic agents that enter the body through the GI tract. Unfortunately, the liver often treats drugs as toxic agents and metabolizes a large fraction before the drug reaches the blood circulatory system; this is called the first-pass effect. For example, in the first pass through the liver, 70–90% of the hormone estradiol is lost. Therefore, when used for hormone replacement therapy, it must be administered as tablets of 1–2 mg/day to achieve effective systemic blood levels. When the same drug is delivered directly to the systemic blood circulation from a transdermal patch, a delivery rate of only 50 μg of estradiol/day is sufficient to achieve the required effect.

II. METHODS OF ACHIEVING CONTROLLED RELEASE

A. Membrane Diffusion-Controlled Systems

In membrane diffusion-controlled systems, a drug is released from a device by permeation from its interior reservoir to the surrounding medium. The rate of diffusion of the drug through the membrane governs its rate of release. The reservoir device illustrated in Fig. 3 is the simplest type of diffusion-controlled system. An inert membrane encloses the drug, which diffuses through the membrane at a finite, controllable rate. If the concentration of the material in equilibrium with the inner surface of the enclosing membrane is constant, then the concentration gradient, that is, the driving force for diffusional release of the drug, is constant as well. This occurs when the inner

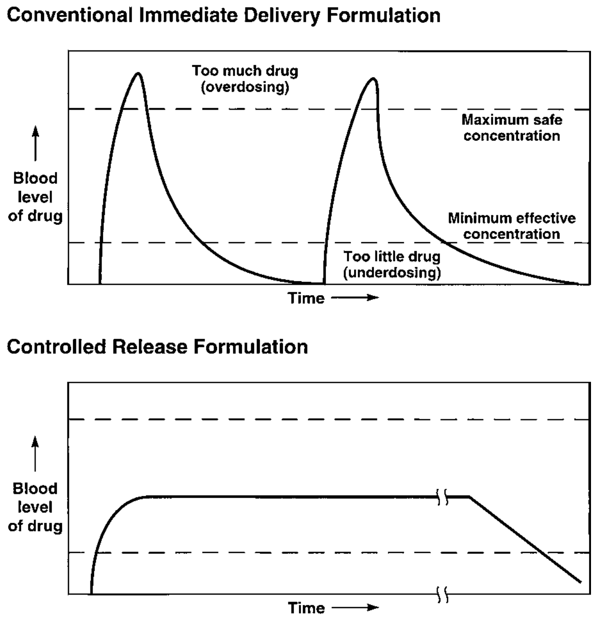


FIGURE 2 Simplified blood level profile illustrating the difference between repeated treatment with a conventional instant delivery formulation and a single, controlled release, long-acting formulation.

reservoir contains a saturated solution of the drug, providing a constant release rate for as long as excess solid is maintained in the solution. This is called zero-order release. If, however, the active drug within the device is initially present as an unsaturated solution, its concentration falls as it is released. The release rate then declines exponentially, producing a first-order release profile. The drug release profile for a simple membrane reservoir system is also shown in Fig. 3.

The drug release rate from a membrane reservoir device depends on the shape of the device, which may be a simple laminate, a cylindrical device, or a spherical device. For the simple laminate, or sandwich geometry, the drug release rate can be written

$$\frac{dM_t}{dt} = \frac{AP\Delta c}{l}, \tag{1}$$

where dM_t/dt is the device release rate, A is the total area of the laminate, P is the membrane permeability, l is the membrane thickness, and Δc is the difference in drug concentration between the solution at the inside surface of the membrane and the drug concentration in the solution at the outer surface of the membrane, usually close to zero. When the solution inside the enclosure is saturated, the drug concentration is c_s, and Eq. (1) reduces to

$$\frac{dM_t}{dt} = \frac{APc_s}{l}. \tag{2}$$

Drug release is then constant as long as a saturated solution is maintained within the enclosure. The total duration of constant release depends on the initial volume of the enclosure V, the mass of encapsulated drug M_0, and the solubility of the drug c_s. The mass of agent that can be delivered at a constant rate is $M_0 - c_sV$. Thus, it follows that the duration of constant release t_∞ is

$$t_\infty = \frac{M_0 - c_sV}{dM_t/dt}. \tag{3}$$

A second type of diffusion-controlled system is a monolithic or matrix device in which the drug is dispersed uniformly throughout the rate-controlling polymeric medium. The drug release rate is then determined by its loading in the matrix, the matrix material, the shape of the device (flat disk, cylinder, or sphere), and the permeability of drug in the matrix material. Equations describing release from all the various device types and geometries can be found elsewhere. As an example, desorption of drug uniformly dissolved in a simple disk (slab)-shaped device can be expressed by either of the two series

$$\frac{M_t}{M_0} = 4\left(\frac{Dt}{l^2}\right)^{1/2}\left[\pi^{-1/2} + 2\sum_{n=0}^{\infty}(-1)^n \operatorname{ierfc}\left(\frac{nl}{2\sqrt{Dt}}\right)\right] \tag{4}$$

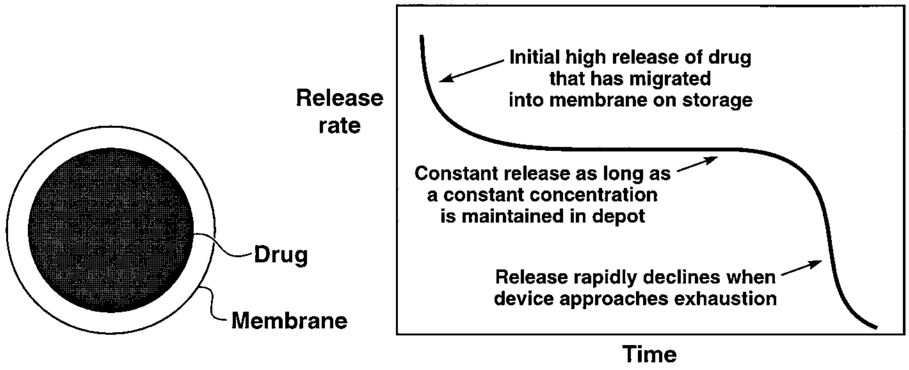


FIGURE 3 Schematic and drug release profile of the simplest form of membrane diffusion-controlled drug delivery system.

or

$$\frac{M_t}{M_0} = 1 - \sum_{n=0}^{\infty} \frac{8\exp[-D(2n+1)^2\pi^2 t/l^2]}{(2n+1)^2\pi^2}, \quad (5)$$

where M_0 is the total amount of drug sorbed, M_t is the amount desorbed at time t, and l is the thickness of the device.

Fortunately, these expressions reduce to two much simpler approximations, reliable to better than 1%, valid for different parts of the desorption curve. The early-time approximation, which holds over the initial portion of the curve, is derived from Eq. (4):

$$\frac{M_t}{M_0} = 4\left(\frac{Dt}{\pi l^2}\right)^{1/2} \quad \text{for} \quad 0 \le \frac{M_t}{M_0} \le 0.6. \quad (6)$$

The late-time approximation, which holds over the final portion of the desorption curve, is derived from Eq. (5):

$$\frac{M_1}{M_0} = 1 - \frac{8}{\pi^2}\exp\left(\frac{-\pi^2 Dt}{l^2}\right) \quad \text{for} \quad 0.4 \le \frac{M_t}{M_0} \le 1.0. \quad (7)$$

The release rate is easily obtained by differentiating Eqs. (6) and (7) to give

$$\frac{dM_t}{dt} = 2M_0\left(\frac{D}{\pi l^2 t}\right)^{1/2} \quad (8)$$

for the early time approximation and

$$\frac{dM_t}{dt} = \frac{8DM_0}{l^2}\exp\left(-\frac{\pi^2 Dt}{l^2}\right) \quad (9)$$

for the late time approximation.

These two approximations are plotted against time in Fig. 4. For simplicity, M_0 and D/l^2 have been set to unity. The release rate falls off in proportion to $t^{-1/2}$ until 60% of the agent has been desorbed, after which the decay is exponential. Although the release rate from monolithic devices is far from constant, this defect is often offset by their ease of manufacture.

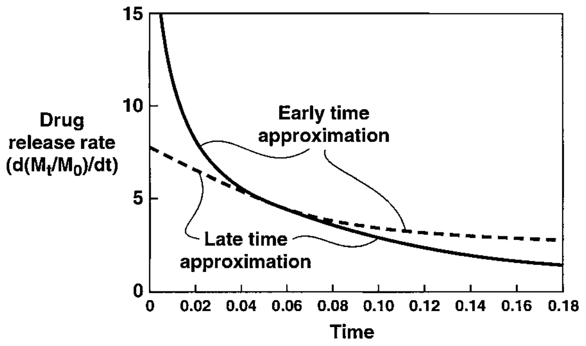


FIGURE 4 Drug release rates as a function of time, showing early- and late-time approximations.

B. Biodegradable Systems

The diffusion-controlled devices outlined above are permanent, in that the membrane or matrix of the device remains in place after its delivery role has been completed. In applications in which the controlled release device is an implant, the drug-depleted device shell must be removed after use. This is undesirable; such applications require a device that degrades during or after completion of its delivery role.

Several polymer-based devices that slowly biodegrade when implanted in the body have been developed. The most important materials are those based on polylactic acid, polyglycolic acid, and their copolymers, as shown in Fig. 5. Other, less widely used biodegradable materials include the poly(ortho esters), polycaprolactone, polyanhydrides, and polycarbonates.

In principle, the release of an active agent can be programmed by dispersing the material within such polymers, with erosion of the polymer effecting release of the agent. One class of biodegradable polymers is *surface eroding*: the surface area of such polymers decreases with time as the cylindrical- or spherical-shaped device erodes. This results in a decreasing release rate unless the geometry of the device is appropriately manipulated or the device is designed to contain a higher concentration of the agent in the interior than in the surface layers. In a more common class of biodegradable polymer, the initial period of degradation occurs slowly. Thereafter, the degradation rate increases rapidly and the bulk of the polymer then erodes in a comparatively short time. In the initial period of exposure to the body, the polymer chains are being cleaved but the molecular weight remains high. Therefore, the mechanical properties of the polymer are not seriously affected. As chain cleavage continues, a point is reached at which the polymer fragments become swollen or soluble in water; at this point the polymer begins to dissolve. This type of polymer can be used to make reservoir or monolithic diffusion-controlled systems that degrade after their delivery role is complete. A final category of polymer has the active agent covalently attached by a labile bond to the backbone of a matrix polymer. When placed at the site of action, the labile bonds slowly degrade, releasing the active agent and forming a soluble polymer. The methods by which these concepts can be formulated into actual practical systems are illustrated in Fig. 6.

C. Osmotic Systems

Yet another class of delivery devices uses osmosis as the driving force. Osmotic effects are often a problem in diffusion-controlled systems because imbibition of water swells the device or dilutes the drug. However, several

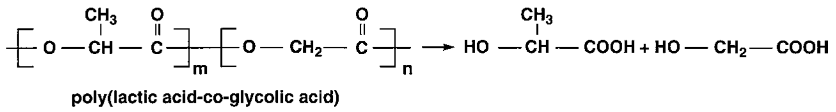


FIGURE 5 Biodegradation of poly(lactic acid-co-glycolic acid) to its simple acid precursors. The rate of biodegradation is a function of the copolymer composition. The pure homopolymers are both relatively crystalline and biodegrade slowly, but the more amorphous copolymers biodegrade more rapidly.

devices that actually use osmotic effects to control the release of drugs have been developed. These devices, called osmotic pumps, use the osmotic pressure developed by diffusion of water across a semipermeable membrane into a salt solution to push a solution of the active agent from the device. Osmotic pumps of various designs are widely applied in the pharmaceutical area, particularly in oral tablet formulations.

The forerunner of modern osmotic devices was the Rose–Nelson pump. Rose and Nelson were two Australian physiologists interested in the delivery of drugs to the gut of sheep and cattle. Their pump, illustrated in Fig. 7, consists of three chambers: a drug chamber, a salt chamber containing excess solid salt, and a water chamber.The drug and water chambers are separated by a rigid, semipermeable membrane. The difference in osmotic pressure across the membrane moves water from the water chamber into the salt chamber. The volume of the salt chamber increases because of this water flow, which distends the latex diaphragm separating the salt and drug chambers, thereby pumping drug out of the device.

The pumping rate of the Rose–Nelson pump is given by the equation

$$\frac{dM_t}{dt} = \frac{dV}{dt} \cdot c, \tag{10}$$

where dM_t/dt is the drug release rate, dV/dt is the volume flow of water into the salt chamber, and c is the concentration of drug in the drug chamber. The osmotic water flow across a membrane is given by the equation

$$\frac{dV}{dt} = \frac{A\theta\Delta\pi}{l}, \tag{11}$$

where dV/dt is a water flow across the membrane of area A, thickness l, and osmotic permeability θ ($cm^3 \cdot cm/cm^2 \cdot hr \cdot atm$), and $\Delta\pi$ is the osmotic pressure difference between the solutions on either side of the membrane. This equation is only strictly true for completely selective

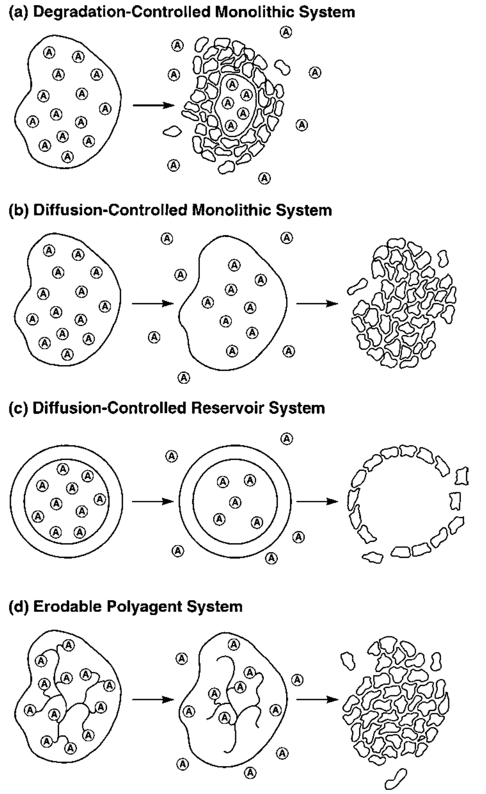


FIGURE 6 Various types of biodegradable drug delivery systems.

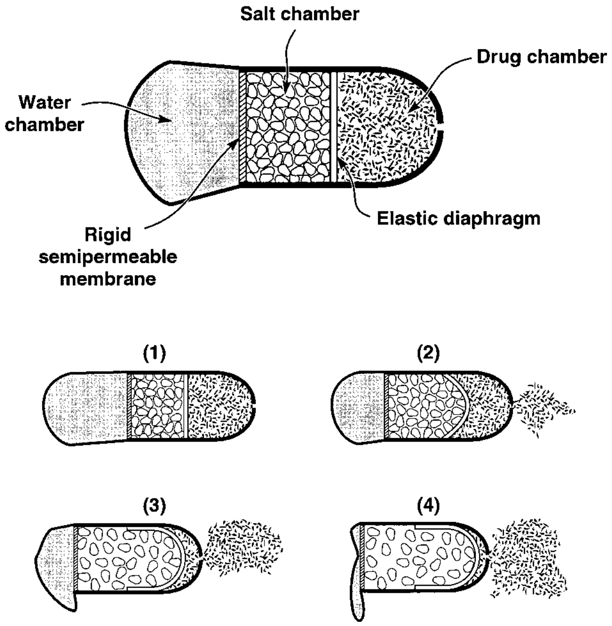


FIGURE 7 Mechanism of action of a Rose–Nelson osmotic pump, the precursor of today's osmotic controlled release delivery systems.

membranes, that is, membranes permeable to water but completely impermeable to the osmotic agent. However, this is a good approximation for most membranes. Substituting Eq. (11) for the flux across the membrane gives

$$\frac{dM_t}{dt} = \frac{A\theta\Delta\pi c}{l}. \qquad (12)$$

The osmotic pressure of the saturated salt solution is high, on the order of tens of atmospheres, and the small pressure required to pump the suspension of active agent is insignificant in comparison. Therefore, the rate of water permeation across the semipermeable membrane remains constant as long as sufficient solid salt is present in the salt chamber to maintain a saturated solution and hence a constant-osmotic-pressure driving force.

Variations of the Rose–Nelson pump have been developed as tools for drug delivery tests in animals. However, the development that made osmotic delivery a major method of achieving controlled drug release was the invention of the elementary osmotic pump by Theeuwes in 1974. The concept behind this invention is illustrated in Fig. 8. The device is a simplification of the Rose–Nelson pump, and eliminates the separate salt chamber by using the drug itself as the osmotic agent. The water required to power the device is supplied by the body, so a separate water chamber is also eliminated. The Theeuwes device is formed by compressing a drug having a suitable osmotic pressure into a tablet using a tableting machine. The tablet is then coated with a semipermeable membrane, usually cellulose acetate, and a small hole is drilled through the membrane coating. When the tablet is placed in an aqueous environment, the osmotic pressure of the soluble drug inside the tablet draws water through the semipermeable coating, forming a saturated aqueous solution inside the device. The membrane does not expand, so the increase in volume caused by the imbibition of water raises the hydrostatic pressure inside the tablet slightly. This pressure is relieved by a flow of saturated agent solution out of the device through the small orifice. Thus, the tablet acts as a small pump, in which water is drawn osmotically into the tablet through the membrane wall and then leaves as a saturated drug solution through the orifice. This process continues at a constant rate until all the solid drug inside the tablet has been dissolved and only a solution-filled shell remains. This residual dissolved drug continues to be delivered, but at a declining rate, until the osmotic pressures inside and outside the tablet are equal. The driving force that draws water into the device is the difference in osmotic pressure between the outside environment and a saturated drug solution. Therefore, the osmotic pressure of the dissolved drug solution has to be relatively high to overcome the osmotic pressure of the body. For drugs with solubilities greater than 5–10 wt%, this device functions very well. Later variations on the simple osmotic tablet design use water-soluble excipients to provide part of the osmotic pressure driving force; this overcomes the solubility limitation.

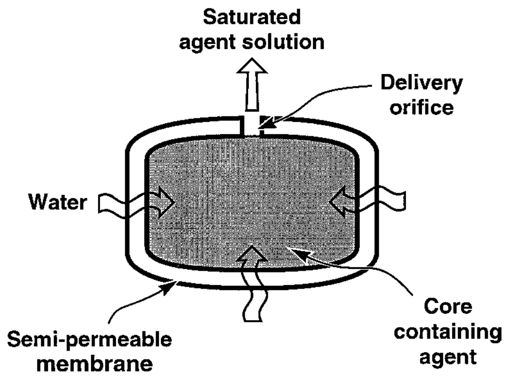


FIGURE 8 The Theeuwes elementary osmotic pump. This simple device consists of a core of water-soluble drug surrounded by a coating of a water-permeable polymer. A hole is drilled through the coating to allow the drug to escape.

III. IMPORTANT CONTROLLED RELEASE PRODUCTS

The controlled release sector is a rapidly expanding part of the pharmaceutical industry. Growth has occurred at the remarkable annual rate of 15% over the past decade, fueled by an explosion of new technologies. The value of the pharmaceuticals using controlled drug delivery reached $20 billion in 1999, and while only modest growth in the overall pharmaceutical market is projected for the next few years, the drug delivery share of the market is expected to continue to grow at least 15% per annum. As much as 20% of the U.S. pharmaceutical market is projected to be controlled release products by 2005.

The majority of drug delivery products reach the market as a result of a strategic alliance between a drug delivery company, which supplies the technology, and a pharmaceutical company, which supplies the drug and the resources needed for full development. A good example of such a collaboration is provided by protein and peptide drugs, an increasingly important area of pharmaceuticals driven by recent advances in biotechnology. Currently, a major factor limiting the use of these drugs is the need for their administration by repeated injections. This is because peptides and proteins are poorly absorbed in the GI tract, so cannot be delivered as oral tablets, and have very short biological half-lives, so cannot be delivered as a single, long-acting injection. Several specialized drug delivery companies are developing innovative techniques that will allow these drugs to be delivered orally, by nasal

inhalation, or in a long-acting injectable formulation. Alliances between these technology providers and the pharmaceutical companies producing the protein/peptide active agents increase the likelihood that these novel drug formulations will reach the marketplace.

Controlled drug delivery has a record of success in the pharmaceutical industry and can offer a high return on capital investment. Some relatively low-budget development programs, which have enabled existing, effective drugs to be administered to the patient by controlled release technology, have been very successful; examples include Lupron Depot (leuprolide, a hypothalamic releasing hormone used for the suppression of testosterone in the treatment of malignant neoplasms of the prostate), Procardia XL (nifedipine, a calcium channel blocker used for the treatment of hypertension and angina), and Cardizem CD (diltiazem, a calcium channel blocker with properties similar to nifedipine). These are all billion-dollar products that employ controlled release or delivery techniques. To develop a new chemical entity through to regulatory approval in the mid-1990s took, on average, 10–12 years and cost about $300–600 million. In contrast, an existing drug can be reformulated into an innovative drug delivery system in 5–7 years at a cost of about $20–$100 million. Interestingly, there are also recent examples in which controlled release is no longer a simple reformulation of an old drug. For example, new drugs are being developed and marketed for the first time as controlled release products. Some of these drugs might not have reached the market except for controlled release technology—felodipine and omeprazole are examples.

A. Oral Formulations

The oral route is by far the most common and convenient method of delivering drugs to the body. Unfortunately, the method has a number of problems that interfere with effective drug delivery. First, a drug taken by mouth is immediately exposed to low-pH stomach acids containing high concentrations of digestive enzymes. Many drugs are chemically degraded or enzymatically metabolized in the stomach before they are absorbed. Drugs that are absorbed then enter the portal circulation and may be destroyed by the first-pass metabolism in the liver described earlier. Controlled release is a method of avoiding these problems.

The typical transit time of material through the GI tract is 12–18 hr, so most controlled release oral formulations are designed to deliver their loading of drug over a 6- to 15-hr period. In this way the action of short-half-life drugs or rapidly absorbed drugs can be spread over a prolonged period. Drugs that might require dosing two or three times a day to achieve relatively uniform and nontoxic blood levels can then be dispensed as a single once-a-day tablet. This simple change, by improving patient compliance and producing a more controlled constant blood level, has produced measurable improvements in efficacy and reduced toxicity for many drugs. Many oral controlled release formulations are designed to produce a relatively low drug delivery rate for the first 1–3 hr while the formulation is in the stomach, followed by prolonged controlled release of the drug once the formulation has reached the GI tract. This avoids chemical degradation of the drug in the aggressive environment of the stomach. This type of delivery is, for example, particularly important for polypeptide drugs which are rapidly and completely destroyed if delivered to the stomach. Delivery to the GI tract is also done to achieve local delivery of the drug, such as the anti-inflammatory Mesalazine for irritable bowel disease and ulcerative colitis.

The precursors of today's controlled release oral formulations were enteric tablets based on various wax matrices designed to circumvent degradation in the stomach. Enteric formulations were later improved by using new, more reliable polymers. By the mid-1970s, the first oral controlled drug delivery systems began to appear. Two important delivery technologies developed at that time were Alza's Oros osmotic controlled release system and Elan's Sodas multiparticulate system. Elan's Sodas system consisted of large numbers of micropellets, each designed to release a microdose of drug by diffusion from a matrix at a predetermined rate. By blending pellets with different release profiles, the overall target rate was achieved. Since then, a wide variety of other oral formulations using osmosis and diffusion have been produced, as well as slow-release bioerodible tablets, ion exchange beads, multiple-layer tablets, and others.

If the drug is relatively water-soluble, osmotic or simple table formulations are often used to achieve controlled delivery. However, with more-insoluble drugs, release of the complete dosage from a single tablet in an 8- to 12-hr period may be difficult. For such drugs, a microencapsulated or granulated form of the drug is enclosed in a gelatin capsule. Microencapsulation exposes a much greater surface area of the device to interact with the body, so drugs that dissolve and diffuse slowly can still be completely released in an 8- to 12-hr period. Drugs can be microencapsulated by physical and chemical methods. Physical methods include encapsulation by pan coating, gravity flow, centrifugation, and fluid bed coating. Chemical microencapsulation normally involves a two-step process called coacervation. Drug particles or droplets of drug solution are first suspended in a polymer solution. Precipitation of the polymer from solution is then caused by, for example, changing the temperature or adding a nonsolvent. The polymer then coats the drug particles to form the microcapsule.

The leading developers of oral drug delivery formulations are Alza and Elan. Other important producers are Skyepharma, which has developed a technology called Geomatrix (a multilayer tablet with each layer releasing the drug at a different rate), R. P. Scherer, which has several different technologies including Zydis (a lyophilized tablet), and Eurand, with several technologies for controlled release, taste masking, and improved bioavailability.

B. Transdermal Systems

Scopolamine, for control of motion sickness, was the first drug to be marketed in the form of a transdermal patch system. Since then the market for transdermal patches has grown steadily. However, the number of drugs that can be delivered through the skin is more limited than was anticipated when the first patches were introduced in the 1980s. The main problem limiting widespread use is the low permeability of most drugs through the skin. Depending on the drug, skin permeabilities are in the range 0.01–10 μg of drug/cm^2 · hr. Because the maximum acceptable size of a transdermal patch is limited to about 50 cm^2, drugs delivered through the skin must be effective at doses of 0.01 mg/day for a poorly skin-permeable drug and 10 mg/day for a highly skin-permeable drug. Very active skin-permeable drugs are required to make transdermal drug delivery possible. Despite the enormous interest in transdermal delivery in academia and industry, the number of drugs delivered transdermally is currently limited to nitroglycerin, nicotine, estradiol, clonidine, fentanyl, testosterone, and isorbide nitrate. Nitroglycerin and nicotine are currently the most important drugs, but the area of hormone replacement therapy is growing as a result of improved estradiol and testosterone delivery patches.

The three main types of diffusion-controlled transdermal devices are shown schematically in Fig. 9. The simple adhesive device (Fig. 9, top) has a two-layer "Band-aid" configuration comprising the backing layer coated with adhesive. The drug is mixed in the adhesive layer used to fix the bandage to the skin. These medicated bandages bring a known quantity of drug to a known area of skin for a known period of time, but have no mechanism for controlling the rate at which the drug is delivered to the patient.

The second type of device is a monolithic system (Fig. 9, middle) incorporating a backing layer, a matrix layer, and an adhesive layer. The matrix layer consists of a polymer material in which the solid drug is dispersed; the rate at which the drug is released from the device is controlled by this polymer matrix. With this type of system, the drug release rate falls off with time as the drug in the skin-contacting side of the matrix is depleted.

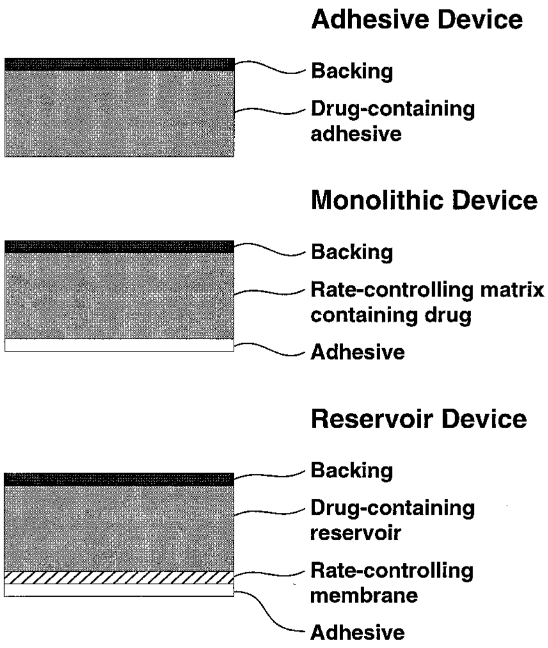


FIGURE 9 Transdermal patch designs.

The third type of device is the reservoir system (Fig. 9, bottom). In this case, the drug—usually in liquid or gel form—is contained in a reservoir separated from the skin by an inert membrane that controls the rate at which drug is delivered to the skin. These devices offer an important advantage over the monolithic geometry: as long as the drug solution in the reservoir remains saturated, the drug release rate through the membrane is constant.

The pattern of drug release from the device is important. If drug is delivered to the skin at less than the maximum rate at which it can be absorbed, the device is the primary dosage-controlling mechanism. When the drug is delivered faster than the skin can absorb it, the skin surface is then saturated with drug at all times, and the limiting factor for systematic dosage is the rate of absorption through the skin. Thus, at least in principle, devices for which the dosage-controlling mechanism is either the skin or the device can be designed.

To reach the systemic blood circulatory system, drug from a transdermal patch must cross several layers of skin, as shown in Fig. 10. The top surface layer of skin, called the *stratum corneum*, represents the main barrier to drug permeation. The stratum corneum is only 10–15 μm thick, but it consists of layers of flattened, cornified cells that are quite impermeable. The interspace between these cells is filled with lipids, giving the structure a "bricks-and-mortar" form, with the cells being the bricks and the lipids being the mortar. The most important pathway for drug absorption is through the lipid (mortar), which dictates

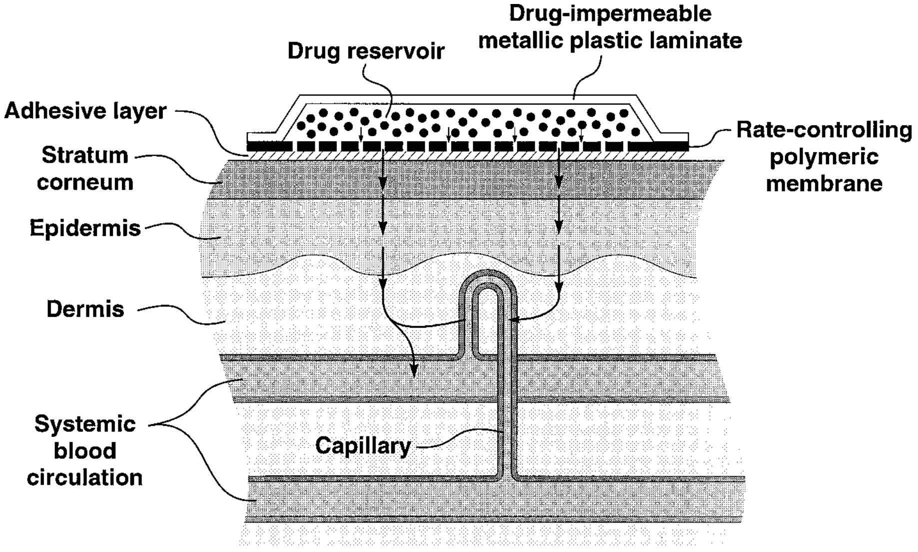


FIGURE 10 Membrane-moderated transdermal drug delivery system (not to scale). Drug permeates from the patch reservoir through the protective outer layers of the skin and is absorbed into the underlying capillaries of the general systemic blood circulation.

the characteristics of usefully permeable drugs. These are drugs with a molecular weight of less than 1000 dalton, a low melting point, an octanol/water partition coefficient between 0 and 2, and few polar centers.

Many drugs do not have the chemistry required to achieve high skin permeabilities, so various methods of enhancing drug permeability have been developed. One method is to soften and swell the stratum corneum by dissolving or dispersing the drug in a simple solvent. For example, the first estradiol-delivery transdermal patch used ethanol to enhance the skin's permeability to estradiol. In the absence of ethanol, estradiol flux through the skin is very low, on the order of 0.01 $\mu g/cm^2 \cdot hr$. However, if a suspension of estradiol in ethanol is applied to the skin, the estradiol flux increases 10- to 20-fold. A second way to enhance skin permeability is to use a compound such as a long-chain fatty acid, ester, or alcohol. These compounds penetrate the stratum corneum more slowly than small molecules such as ethanol but have a more prolonged plasticizing effect. A third method combines elements of the first two. An enhancer formulation containing both a simple solvent and a fatty component is used to combine the rapid onset of solvent effect with the prolonged action of the fatty component.

Another method of enhancing drug permeation is to increase the driving force for drug permeation by using a small electric current. This last approach, called iontophoresis, has been widely studied. The principle of iontophoresis is illustrated in Fig. 11. In this method, a battery is connected to two electrodes on the skin. An ionized drug placed in contact with one electrode will migrate under the influence of the voltage gradient through the skin and enter the system circulation; very large enhancements can be obtained. The earliest devices having the essential features of iontophoresis date back to the 1890s, although apparently their objective was to shock their subjects rather than to administer drugs to them. The first modern device appeared in 1972, and advances in electronics have since allowed smaller and smaller devices to be built. The newest devices have a built-in battery layer and are comparable in size to a normal transdermal patch. Iontophoresis appears to be a particularly promising tool for the delivery of very active peptides, small proteins, or oligonucleotides, which are otherwise almost completely

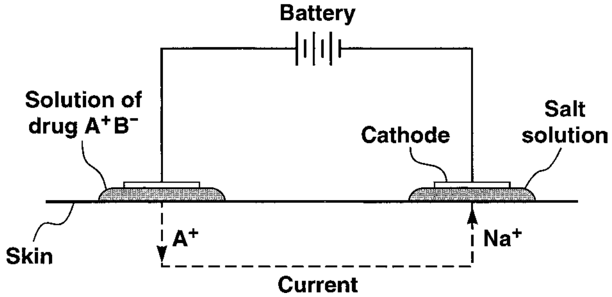


FIGURE 11 Mechanism of an iontophoresis patch. The poles of a small battery are connected to the skin. A solution of an ionized drug placed at one electrode migrates through the skin into the systemic circulation under the action of the voltage driving force of the battery.

skin-impermeable. Under the action of a small voltage gradient, the skin permeation rate of the drug increases 10- to 100-fold. Another advantage of iontophoresis is that the voltage driving force can be easily regulated; this allows pulsatile drug delivery or variable delivery according to the patient's needs. This control is useful in the delivery of analgesics such as fentanyl for the control of pain. Despite the many patents in this field, no commercial products have yet reached the market.

The main advantages of transdermal over oral drug delivery are the ability to maintain constant plasma levels with short-half-life drugs and to avoid the hostile conditions of the gastrointestinal tract and consequent drug deactivation because of the hepatic first-pass effect. Moreover, patient compliance tends to increase, reducing the number of administrations required. Patches are now available for once-a-day, twice-a-week, and once-a-week treatment.

C. Nasal Spray/Inhalers

Drug delivery by inhalation has a long history and is an obvious method of administering agents that act on the respiratory system. However, it is now being used increasingly to deliver systemically active drugs. The first hand-held, pressurized, metered inhaler was launched in 1960; several other products have been introduced since. Intranasal delivery is currently employed in treatments for migraine, smoking cessation, acute pain relief, nocturnal enuresis, osteoporosis, and vitamin B_{12} deficiency. In 1999, Aviron's intranasal influenza vaccine, FluMist, was filed with the FDA, and several other vaccines are being considered for nasal administration. Other applications for nasal delivery under development include cancer therapy, epilepsy control, antiemetics, and treatment of insulin-dependent diabetes.

The advantages of the nasal cavity over other drug administration routes are rapid, direct access to the systemic circulation and complete avoidance of first-pass metabolism. The nasal cavity is also a far less aggressive environment than the gastrointestinal tract, and so is particularly useful for the delivery of peptides and proteins, which are easily degraded in the stomach. Patient compliance is also improved, particularly if the alternative treatment is intravenous injection. However, the nasal route is not without its problems. Issues that need to be considered are the rate of drug absorption through the nasal mucosa, the residence time of the formulation at the site of delivery, local toxicity and tolerability, and degradation of the drug in the nasal cavity. The rate of absorption of the drug can be controlled by including carriers and enhancers to modify permeation, or by chemical modification of the drug into a more readily absorbed form. Most nasal devices deliver the drug as a fine liquid or solid spray; the average particle size significantly affects the rate of drug absorption. Intranasal delivery of systemic drugs requires sophisticated delivery devices to ensure accurate, repeatable dosing and minimal irritation. Dosing accuracy is particularly important for delivery of potent and expensive drugs such as the new generation of peptide and protein products. Long-term nasal delivery may also require dose-counters and lock-out systems to prevent overdosing as, for example, in the pain relief area.

Prior to 1985, chlorofluorocarbons were widely used as inert nontoxic inhaler propellants; these compounds are now all but banned. Consequently, much recent research has centered on the development of dry powder inhalers, which deliver the active ingredient as ultrafine particles directly to the lungs with minimal deposition in the mouth and trachea. Some of these devices are activated directly by the inspiration of the patient without the need to coordinate activation and inspiration.

Although dry powder inhalers have a number of therapeutic benefits, they also have problems. For example, contact of the powder formulation with moisture can lead to agglomeration and inaccurate dosing, and dose uniformity is hard to achieve. The potential for nasal delivery is greatest in two areas: local delivery to the lung for respiratory treatment diseases and systemic delivery of a broad variety of drugs via the alveoli, which have highly absorptive properties. Its greater long-term potential is in the delivery of macromolecules, but further research is needed to determine the long-term immunogenicity, reproducibility, and stability of the delivery systems.

D. Targeted Drug Delivery

Targeted or directed drug delivery is a relatively new method of delivering drugs. The objective is to alter the effectiveness of drugs by targeting their delivery directly to the site needing drug action. Promising techniques include the use of liposomes, polyethylene glycol (PEG)-coated molecules, blood–brain barrier transfer agents, and several antibody conjugate approaches.

The most advanced targeted delivery technology uses liposomes, which are ultrafine water/oil/water emulsions in which the emulsion droplets consist of lipid vesicles containing an aqueous drug solution. The surface of the vesicles is sometimes modified to promote targeting of the lipid-drug-containing vesicle to selected tissues. The first liposomes were developed in 1965 as a model of biological membranes. Their potential as a drug delivery system was recognized later; now, after many years of gestation, liposomes are finally being introduced in commercial

products. The mean diameter of liposomes is less than 0.1 μm. This allows them selectively to extravasate into tissues characterized by leaky vasculature, such as solid tumors, achieving targeted delivery to the diseased organ with low adverse effects on normal tissues.

The first liposome product on the market was Ambisome, containing amphotericin B, an antifungal encapsulated in liposomes to reduce its toxicity. Since then, liposomal preparations of other drugs have been developed, most importantly Daunoxome, an anticancer product containing the antitumoral drug daunorubicin for the treatment of Kaposi's sarcoma. Liposomes are also being investigated as vehicles for gene therapy. Their main advantages over viral carriers is a higher loading capacity and a lower risk of evoking an immune response. Surface-coating liposomes with antibodies to allow active targeting to a particular disease site is also being studied in clinical trials. If these products are successful, targeted drug delivery will be a breakthrough in the treatment of a number of major diseases.

A second approach to targeted drug delivery uses polyethylene glycol (PEG), a water-soluble polymer covalently linked to proteins, which alters their properties and extends their potential use. The main product using this approach is for α-interferon for the treatment of hepatitis C. This technique has been applied to several proteins including adenosine deaminase, cytokines, and granulocyte-macrophage colony-stimulating factor. Generally the results obtained by the modification with PEG are increased circulating life, reduced immunogenicity and antigenicity, and increased stability and solubility with a minimal loss of biological activity.

E. Implants

Three types of implantable drug delivery devices are currently in use or being developed. The first type is an implantable polymeric capsule most commonly made from silicone rubber, which when placed under the skin delivers the drug load at a constant rate for as long as 2–5 years. Upjohn's Depo-Provera and American Homes' Norplant, used to deliver contraceptive steroids for long-term contraception, are examples of this approach. Implantable devices achieve long-term, controlled delivery of the drug and patient compliance is no longer an issue, both significant advantages. The key disadvantage is the size of the device, which means minor surgery is required to insert the implant and later to remove the device once its drug delivery role has been completed. There is also a potential to release all of the drug in a few days if the capsule begins to leak. For these reasons this type of nondegradable device has not become a major product.

A second type of device uses implants made from biodegradable polymers. Because the device is biodegradable, device retrieval once drug delivery is complete is no longer necessary. This makes the device much more acceptable to patients. Unfortunately, it is technically difficult to create implantable, biodegradable devices that deliver drug reliably for more than 1 or 2 months. Therefore, these devices are generally restricted to delivering drugs for 4–6 weeks. Examples include Abbot's Lupron Depot formulation, a single monthly injection of leuprolide for endometriosis, and Zoladex, Zenaca's 4-week implantable formulations of goserelin for endometriosis in women and prostrate cancer in men.

A final category of implantable devices, in late-stage development, is miniature pumps driven by osmosis or fluorocarbon propellants. These pumps are designed to deliver essentially any drug at a predetermined rate for weeks or even months. The problem to be solved in this case is to make the pump small enough and reliable enough that the trauma of device removal is outweighted by its drug delivery benefits.

IV. FUTURE DIRECTIONS

The era of modern controlled drug delivery started with the launching of the first product registered at the U.S. Food and Drug Agency in terms of both the total amount of drug delivered and the rate of drug delivery. This first product, Alza's Ocusert, was launched in 1974. By 1990 the total controlled release market was approximately \$1 billion. Since then the market has grown 20-fold, and the number of such products is expected to continue to grow rapidly in the next few years.

The largest growth area will be the extension of already developed controlled release technologies to a wider number of drugs. A long list of oral and some transdermal controlled release products are in late-stage clinical trials and should appear on the market in the next few years.

A second area of significant future growth will be the use of controlled release systems to deliver the new generation of peptide and protein drugs. As a consequence of the sequencing of the human genome, the information required to design protein and peptide drugs to treat important diseases is at hand. However, many of these new drugs are essentially ineffective if administered as conventional pharmaceutical formulations. New, well-designed controlled release systems that can deliver the drugs intact at a prolonged steady rate close to the site of action are required; such systems are being actively developed.

A third potential growth area for controlled release technology is the development of systems for targeted drug delivery. Drug delivery to the ultimate site of action is a multistep process. Conventional controlled drug delivery systems generally only address the first step, the rate of delivery of drug to the body. The path of the drug from the dosage site to the diseased cells is largely uncontrolled. Targeted drug delivery systems attempt to improve this step in the delivery process, in effect, to improve the aim of Ehrlich's magic bullet. Antibody-coated liposomes and the PEG-peptide products described above are the first simple prototype product designs to tackle this problem. Over time, more sophisticated and more effective techniques will be developed.

Finally, all the controlled release products described in this article are preprogrammed, with the rate of delivery being established by the designer of the device. No subsequent adjustment of the drug delivery rate in response to patient need is possible. However, many disease conditions are episodic, for example, diabetes, stroke, migraines, heart attacks, and epilepsy. Controlled drug delivery systems that could sense the body's need and automatically deliver the drug at the rate required would be a huge benefit. In the intensive care facilities of modern hospitals, this type of control is achieved through continuous electronic sensors and monitoring by the attending nurses and physicians. In the future, development of microelectronic/micromechanical patient-portable machines to produce the same level of control on ambulatory, at-risk patients can be imagined.

SEE ALSO THE FOLLOWING ARTICLES

Aerosols • Biopolymers • Electrophoresis • Ion Transport Across Biological Membranes • Membranes, Synthetic, Applications • Pharmaceuticals • Pharmacokinetics

BIBLIOGRAPHY

Baker, R. (1987). "Controlled Release of Biologically Active Agents," Wiley, New York.

Benita, S. (1996). "Microencapsulation: Methods and Industrial Applications," Marcel Dekker, New York.

Chasin, M., and Langer, R. (1990). "Biodegradable Polymers as Drug Delivery Systems," Marcel Dekker, New York.

Chien, Y., Su, K., and Chang, S. (1989). "Nasal Systemic Drug Delivery," Marcel Dekker, New York.

Clark, A. (1995). "Medical aerosol inhalers: Past, present and future," *Aerosol Sci. Technol.* **22,** 374–391.

Friend, D. (1992). "Oral Colon-Specific Drug Delivery," CRC Press, Boca Raton, Fl.

Katre, N. (1993). "The conjugation of proteins with polyethylene glycol and other polymers," *Adv. Drug Deliv. Rev.* **10,** 91–114.

Lasic, D., and Papahadjopoulos, D. (1998). "Medical Applications of Liposomes," Elsevier Science, New York.

Potts, R., and Guy, R. (1997). "Mechanisms of Transdermal Drug Delivery," Marcel Dekker, New York.

Santus, G., and Baker, R. (1995). "Osmotic drug delivery: A review of the patent literature," *J. Controlled Release* **35,** 1–21.

Smith, E., and Maibach, H. (1995). "Percutaneous Penetration Enhancers," CRC Press, Boca Raton, FL.

Wise, D. (2000). "Handbook of Pharmaceutical Controlled Release Technology," Marcel Dekker, New York.

Pharmacokinetics

Michael F. Flessner
University of Rochester

I. Basic Pharmacokinetic Concepts
II. Drug Absorption
III. Drug Distribution
IV. Drug Metabolism and Elimination
V. Pharmacokinetic Variability
VI. Quantitative Approaches to Pharmacokinetic Modeling

GLOSSARY

Active transport Typically drug transport via a specific transporter, requiring energy, and moving solute against its electrochemical gradient.

Bioavailability Fraction of drug absorbed unchanged into the target compartment.

Drug clearance Volume of distribution cleared of drug per unit time.

Half-life Time required for the concentration of a drug to decrease by one half ($t_{1/2}$).

Passive transport mechanisms Mechanisms that use the forces of concentration differences (diffusion) or pressure differences (convection) to move substances from one site to another.

Pharmacodynamics The study of the specific cellular and molecular action of a drug on a biological organism.

Pharmacokinetics The science of quantitation of drug absorption, distribution, and elimination within biological organisms. It is a subset of pharmacology.

Pharmacology The study of drugs, their sources, different preparations, and therapeutic uses.

Volume of distribution Apparent volume to which a dose of drug distributes at the time of injection or administration.

PHARMACOKINETICS is the science of quantitation of drug absorption, distribution, and elimination within biological organisms. It is a subset of *pharmacology*, which is the study of drugs, their sources, different preparations, and therapeutic uses. On the other hand, *pharmacodynamics* is the study of the specific cellular and molecular actions of a drug on a biological organism. Figure 1 illustrates these definitions for the general case. Together, pharmacokinetics and pharmacodynamics encompass the entire process of drug entry into the body of the organism to the ultimate biological effect. This article will focus on pharmacokinetics, which applies mathematics to the following questions:

1. How quickly and how much of a drug enters the organism?
2. To where does the drug distribute?
3. How rapidly is the drug removed from the body?

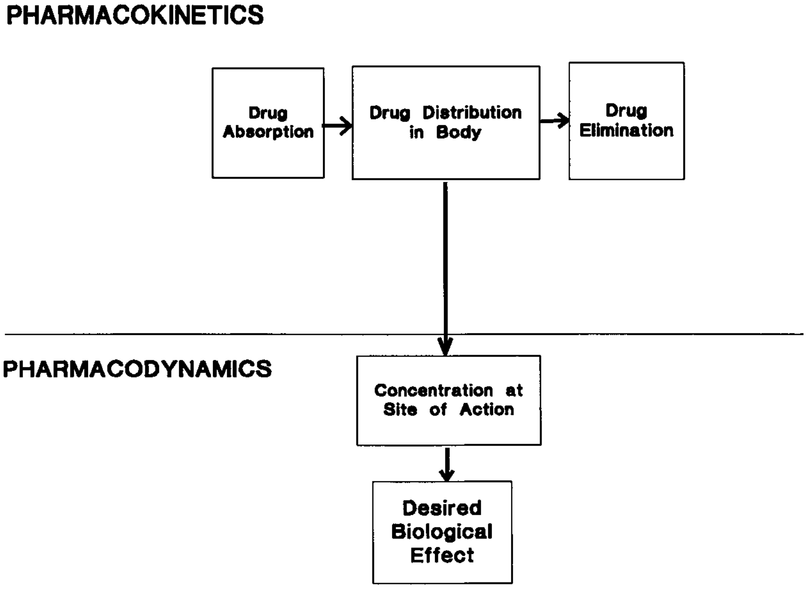


FIGURE 1 Pharmacokinetics versus pharmacodynamics. Pharmacokinetics deals with the phases of drug absorption, drug distribution, and drug elimination. Pharmacodynamics defines the biological effect of a drug at a site of action.

The practical use of the mathematical approaches of pharmacokinetics requires data consisting of measurements of the rate of drug entry into the organism, time-dependent drug concentration in specific compartments within the organism, and the rate of drug removal from the organism. Since the largest amount of data is available for mammals—in particular, humans—mammalian systems will be emphasized in this article. Figure 2 illustrates a more detailed conceptual model of pharmacokinetic processes in humans. Each of these processes will be discussed in turn and specific mathematical approaches will be given. For more detailed approaches, the reader is referred to the Bibliography at the end of the chapter.

I. BASIC PHARMACOKINETIC CONCEPTS

The following basic terminology is used in pharmacokinetics: half-life, volume of distribution, bioavailability, and mechanisms of transport.

A. Half-Life

The *half-life* ($t_{1/2}$) of a drug is equal to the time required for its concentration to decrease by one half. Many drugs follow "first-order" kinetics and, as illustrated in Fig. 3, give a straight line on a semilog plot. This can be mathematically described by the following equation:

$$C = C_0 \exp(-k_{\text{elim}} t), \qquad \text{where} \quad k_{\text{elim}} = \frac{0.693}{t_{1/2}}. \quad (1)$$

Here C_0 is the concentration in the body compartment (typically equal to plasma concentration in a human or animal subject) at time zero, t is the time, and the drug elimination rate constant k_{elim} can be found from the half-life or from the negative slope of the terminal phase of the semilog curve, as illustrated in Fig. 3. The definition

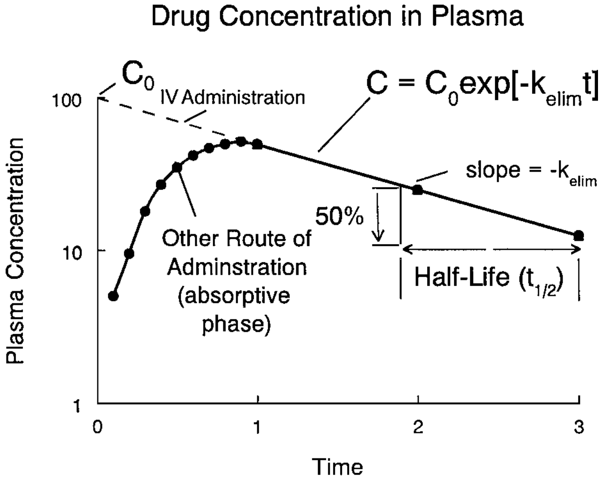


FIGURE 3 Semilogarithmic plot of plasma concentration of a drug versus time in a one-compartment system. The initial part of the curve depends on the route of administration. The half-life of a drug equals the time required for the plasma concentration to decrease by 50%. The drug elimination rate k_{elim} can be found from the terminal slope of the curve. The concentration in the plasma at time zero C_0 is found by extrapolating the plasma concentration curve back to $x = 0$.

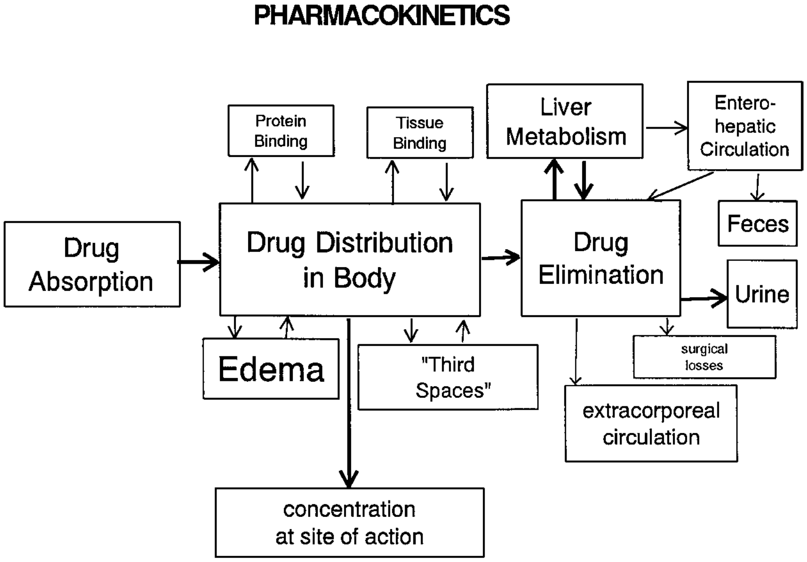


FIGURE 2 Pharmacokinetics: an expanded definition of the three phases of drug delivery to the site of action.

for $t_{1/2}$ is the same for drugs which have more complex curves of concentration versus time or are administered via a route other than the intravenous (i.v.) route. However, the method of calculation may differ from the relatively simple arrangement of Eq. (1). Various texts listed in the Bibliography give more complex formulations.

Drug clearance Clr is analogous to creatinine clearance and is equal to the volume of distribution V_d cleared of drug per unit time:

$$Clr = k_{elim} V_d. \qquad (2)$$

B. Volume of Distribution

The apparent *volume of distribution* V_d is equal to the dose given in an i.v. bolus divided by the plasma (or blood) concentration at time zero, C_0. The value of C_0 can be found by the extrapolation of the plasma concentration curve back to time zero, as shown in Fig. 3. An alternate method of calculation is

$$V_d = \frac{\text{i.v dose (or dose absorbed)}}{\text{area under plasma curve} \times k_{elim}}. \qquad (3)$$

The apparent volume of distribution does not necessarily correspond to any anatomic compartment in the body but is a "virtual" compartment which is mathematically defined by equations such as (3). For example, if a drug such as morphine is bound to protein and tissue, the volume of distribution is several times the total body water volume, which is approximately 60% of the body weight (for morphine $V_d = 3.5$ L/kg of body weight or $\sim$6 L/L body water).

C. Bioavailability

The *bioavailability F* of a drug is equal to the fraction of drug which absorbed unchanged into the target compartment. For example, drugs which are injected intravenously into mammals have a bioavailability of 1 because all of the drug arrives in the plasma, which distributes it to other parts of the body. If other routes are used to administer the drug, such as the oral route, F will be determined by its rate and extent of absorption from the stomach and intestines. Specific routes of absorption will be discussed below.

D. Mechanisms of Transport

Drugs pass across barriers by a number of mechanisms. *Passive mechanisms* utilize the forces of concentration differences or pressure differences to move substances from one site to another. *Active transport* of a drug is typically via a specific transporter, requires energy, and moves solute against its electrochemical gradient. Figure 4 illustrates these definitions.

Most drugs cross membranes via the passive processes of diffusion or convection. *Diffusion* is the process by which a substance moves from a region of high concentration to one of low concentration and is important for movement of low-molecular weight substances (MW $<$ 6000 Da). The governing mathematical expression is

$$\text{rate of diffusion} = -DA\frac{dC}{dx}. \qquad (4)$$

In this equation D is the diffusion coefficient, which depends on the molecular size of the substance, its lipid solubility, the ionic charge, and the characteristics of the

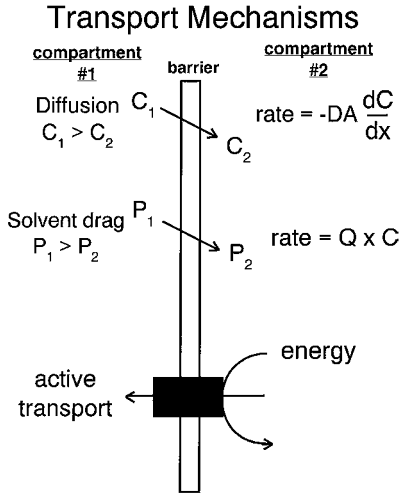


FIGURE 4 Transport mechanisms across biological barriers. Most drugs *diffuse* down their concentration gradient at a rate defined by the product of the diffusion coefficient *D*, the area *A*, and the concentration gradient dC/dx. A second passive mechanism is *convection or solvent drag*, which moves a solute at the effective solvent flow rate *Q* according to the existing pressure gradient. *Active transport* requires energy to move solutes in a direction opposite to the electrochemical gradient.

barrier; A is the area of barrier exposed to the drug; and dC/dx is the concentration gradient within the barrier. Often, it is impossible to define A or the relationship between C and the distance x within the barrier, and the transport is described by the expression

$$\text{rate or mass transfer} = PA(C_1 - C_2), \quad (5)$$

where *PA* is the mass transfer–area coefficient, which lumps the effective area and apparent permeability of the barrier to the solute; C_i is the concentration in each compartment. Often the concentration in the "receiving compartment" or C_2 is unknown, and *PA* is termed a clearance term, defined by the rate of solute loss from the originating compartment to the receiving compartment divided by C_1.

A special form of passive transport is *carrier-mediated transport*, which is important in the handling of urea by the kidney and in the absorption of certain solutes from the intestine. The general expression for carrier-mediated transport is defined by the typical equation for enzyme kinetics

$$\text{transport rate or reaction rate} = \frac{V_{\max} C}{K_M + C}, \quad (6)$$

where $V_{\max}$ is the maximal rate of transport or chemical reaction, K_M is the concentration at which the rate is one-half maximum, and C is the concentration.

Compartments in an organism consist of volumes of water in which substances are in solution. Compartments are separated by the cell membrane, which is typically a lipid bilayer with low permeability to water or water-soluble molecules. Lipid-soluble molecules will therefore transport across such barriers more rapidly than substances which have very poor solubility in lipid. The relative solubility of a substance is often defined as the *partition coefficient*, which is the ratio of the solute's concentration in octanol to its concentration in a water-based buffer. The stomach mucosa and the epithelium of the colon display the property of increasing transport rates with increasing lipid solubility. However, mammalian epithelia such as the rat intestine demonstrate a maximal rate of transport at a partition coefficient between 10 and 20 because of an unstirred fluid layer which is maintained by the microvilli coating the epithelia. In human skin, the partition coefficient for optimal rates of transport appears to be between 200 and 250.

Convection is a process by which a substance is dragged along by the flow of fluid; hence the term *solvent drag* is used to describe this type of transport. The flow is powered by osmotic or hydrostatic pressure gradients which exist across tissue boundaries. The kidney is an example of an organ which depends on hydrostatic pressure-driven convection for filtration of substances by the glomerulus and osmotic-pressure driven convection for solute reabsorption in the proximal tubule. Filtration and reabsorption by blood capillaries depends on Starling's relationship:

$$\text{fluid flux} = L_p\left[(P_1 - P_2) - \sum_i \sigma_i(\pi_{i,1} - \pi_{i,2})\right], \quad (7)$$

where L_p is the hydraulic conductivity of the barrier; P_j is the hydrostatic pressure in compartment j; σ_i is the reflection coefficient of solute i; and $\pi_{i,j}$ is the osmotic pressure exerted by solute i in compartment j. A solute with $\sigma_i = 1$ does not cross the barrier and exerts its full osmotic force on the barrier. If $\sigma_i = 0$, the solute exerts no osmotic force across the barrier.

Molecular size has a significant effect on the rate of transport across biological barriers. Smaller substances such as urea or glucose transport across membranes rapidly via carriers or pores. While substances with molecular weights greater than 1000 Da transport easily across endothelia of most tissues, they are typically absorbed across epithelia very slowly. In the absorption of drugs, proteins, or peptides, pinocytosis or phagocytosis may play a role: *Pinocytosis* is the process by which the cell engulfs material in the fluid external to the cell and transports it into the cell, where it may combine with a lysosome and be metabolized; *phagocytosis* is the process by which resident macrophages and leukocytes take up and digest solid material or particles.

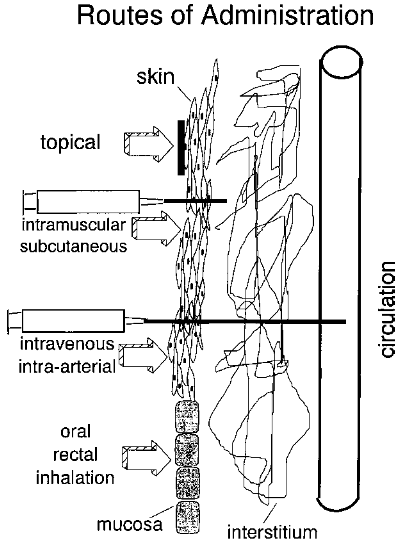


FIGURE 5 Routes of administration; see text for a detailed discussion of each route.

II. DRUG ABSORPTION

Absorption of a drug consists in the movement of the substance from the site of administration to a body compartment from which the drug has access to the target. Routes of administration (see Fig. 5) vary depending on whether a local or systemic effect is desired. For example, a *local effect* may be elicited by application of a drug to the skin or a mucous membrane. The drug must have specific characteristics which limit its movement from the site of administration into the local target but not into the systemic circulation. Administration into the gastrointestinal tract (sublingual, oral, or rectal route), inhalation into the lungs, or injection into the subcutaneous space, the muscle (intramuscular), or the cerebral spinal fluid (intrathecal) are examples of routes which will typically lead to systemic effects. A substance is considered to be absorbed once it arrives in the blood circulation. Drugs administered orally must transport across both epithelial and endothelial barriers to reach the circulation; they may undergo binding within the tissue or uptake into cells and metabolism prior to arrival at the blood capillary wall. The amount of a drug absorbed systemically therefore depends on properties of the drug and the characteristics of the barriers between the site of administration and the circulation.

A. Gastrointestinal Absorption

The gastrointestinal tract includes the oral cavity, the stomach, the small intestines, the large intestine, and the rectum. The entire system is lined with epithelial cells, which typically have tight junctions between them which are relatively impermeable to the passive transfer of solutes and water. Throughout the tract, there is a varying environment of pH and composition.

In the oral cavity, the epithelium is relatively thin and very well vascularized, so that it is potentially an excellent site for absorption of weak electrolytes which would be present in an undissociated form. The epithelium remains moist with saliva, which maintains the pH in a slightly acidic condition. However, the constant secretion and flow of saliva tends to dilute any solution in the oral cavity and move it toward the esophagus. Solid tablets can be held under the tongue, and substances are rapidly absorbed from the tablet into the blood. Nitroglycerin is an example of a drug often administered via this route; it provides rapid relief from chest pain due to poor perfusion of the heart. Advantages of *sublingual administration* include direct absorption into the general circulation, which does not pass through the liver as it would if the drug were absorbed in the stomach or intestines.

The stomach is the next major portal of entry of substances taken orally and swallowed. The rapid transit through the mouth and esophagus leaves little time for absorption in those regions. The makeup of the pill or capsule and how it is taken also determine the site of absorption. If a tablet is taken with water on an empty stomach, it will likely pass rapidly through the stomach and into the small intestine. On the other hand, if taken with food, the substance will likely spend more time in the stomach. The surface of the stomach, called the *gastric mucosa*, is made up of a series of infoldings which greatly increases the absorptive surface area. The pH of the stomach is usually quite low ($pH < 3$ in humans) and therefore substances which are weak acids remain in undissociated form and are easily absorbed. Their rate of absorption will depend on the lipid solubility.

The small intestine has extensive infoldings called villi; the villi are in turn lined by a brush border of microvilli. This hollow organ therefore presents a large absorptive surface for nutrients as well as for drugs. The passage of a substance through the small intestine normally takes several hours and absorption is usually complete. If the small intestine has been shortened through surgery or if the transit time has been decreased by diarrhea, then absorption may not be complete. The major mechanism of transport is diffusion across the epithelium and subsequently across the endothelium of vessels coursing through the tissue. Drugs which are absorbed in this fashion enter the intestinal circulation which flows to the liver via the portal vein. Therefore, drugs absorbed into the circulation of the hollow viscera pass through the liver prior to reaching other organs. This event, which may result in metabolism of the drug, is designated a "first-pass effect." The first-pass effect may activate a drug, form active

intermediates, or result in metabolites with no biological activity.

The large intestine or colon is not designed as a primary site of absorption. It does not have the infoldings or microvilli which the small intestine has. However, the last part of the large intestine, the rectum, is an excellent site of absorption for several reasons. Drugs can be directly introduced via the anal opening into the rectum without passage through the proximal portions of the intestines. Once absorbed into the circulation of the rectum, substances do not undergo a first-pass effect since the blood does not flow through the portal vein but circulates directly to the general circulation.

B. Transdermal, Subcutaneous, Intramuscular Routes of Absorption

The outer layer of the skin, termed the epidermis, presents a major barrier to the passage of water or solute. The outermost, horny layer is made up of dead cell product and keratin and prevents loss of water from the body. Because of the dense cell packing, passive diffusion is slow, but substances are capable of transporting through the skin over a prolonged time. Heavy metals such as mercury or lead or organic solvents or toxic insecticides are absorbed in this fashion. In recent years, transdermal delivery devices have been developed which provide continuous slow transfer of drug from a flat patch reservoir placed directly on the skin. The controlled delivery rate depends on an intact epidermis and dermis and a stable local blood supply. Various studies have demonstrated that drugs are often not totally taken up in the dermis but penetrate to underlying tissue and result in very high local tissue concentrations. These devices are commonly used in the 24-hr delivery of nitrates to patients with heart disease or in the continuous administration of antihypertensive medication to individuals with high blood pressure. Because of the slow infusion properties of these devices, drugs with relatively short half-lives can be administered continuously for up to 1 week. Nitropaste (a paste form of nitroglycerin), for example, can be applied in the form of topical ointment, and the rate of delivery can be adjusted by changing the area of application until intravenous access can be obtained. Some drugs such as nitrates can permeate the skin rapidly; the amount prescribed is typically written in terms of surface area covered. The available surface area may become important in the cardiac patient who has unstable angina and does not have intravenous access. Advantages of the transdermal route include avoidance of hepatic first-pass metabolism, ability to discontinue treatment by removal of the system, ability to continue delivery of the drug for a longer period of time than the usual oral form of the drug, and a potentially large area of application.

Subcutaneous injections deliver the drug just below the dermis and therefore bypass the barrier of the skin. The drug will quickly diffuse to lymphatic or blood capillaries and be taken up and distributed systemically. Because lymphatics take up large molecules, this is an excellent site for injections of high-molecular weight agents such as vaccines.

Intramuscular (i.m.) injections are an alternative when i.v. access is not available. For example, lidocaine has been shown to be rapidly absorbed after i.m. injection, with the deltoid muscle being a superior site to either the buttocks or lateral thigh. Many drugs show great variability in absorption after intramuscular or subcutaneous injection. Because of the uncertainty in the bioavailability of the drug from these routes, in an emergency the i.v. route is clearly superior for rapid administration of a defined dose.

C. Intraperitoneal Administration

The *intraperitoneal* (i.p.) route of administration is particularly advantageous if access to the cavity already exists (i.e., a peritoneal dialysis catheter) and if the infection or disease process is confined to the cavity. Medication levels in the i.p. solution can be maintained one to two orders of magnitude above toxic levels in the plasma. Transfer to the systemic circulation is relatively slow, and peak concentrations in the plasma will be a fraction of what they would be if the same dose were administered i.v. This provides for maximal pharmacologic activity locally while sparing the remainder of the body from the toxic side effects. The pharmacokinetic advantage of regional administration into the peritoneal cavity (or any body cavity) over that of intravenous administration can be calculated by

$$\text{pharmacokinetic advantage} = \frac{(C_{\text{PC}}/C_{\text{B}})_{\text{IP}}}{(C_{\text{PC}}/C_{\text{B}})_{\text{IV}}} = 1 + \frac{Clr_{\text{total body}}}{PA}, \tag{8}$$

where C_{PC} is the concentration in the peritoneal cavity, C_{B} is the concentration in blood, and PA is the mass transfer area coefficient governing transport between the peritoneal cavity and the blood: the mass transfer rate is $PA_{\text{peritoneum}}(C_{\text{PC}} - C_{\text{B}})$.

D. Inhalation

The lungs are the optimum portal of entry of gases and very small particulate aerosols ($<2\mu$m in size). The area of the alveoli is approximately 100–200 m^2, with a very thin (1–2 μm) series of membranes separating the air and blood spaces. Drugs reaching the alveoli are therefore absorbed

very rapidly into the circulation and bypass the first-pass effect of the liver. Aerosols with sizes greater than 10 μm are often trapped in the nasal passages or the pharynx and upper airway. Therapeutic aerosols, such as those used for acute obstructive asthma, need to be designed to deliver particles less than 10 μm. There is a clear advantage of direct delivery to the lungs in the treatment of pulmonary disease in that many side effects of the drugs which might occur if they were administered systemically can be avoided.

E. Mathematical Approaches to the Description of Absorption and Bioavailability

The definition of bioavailability given above includes elements of the rate of drug entry into the target compartment as well as the total amount. Let us assume that the blood circulation is the target compartment and that i.v. administration is the standard route to which all others must be compared. Then F may be calculated as the ratio of the area under the plasma concentration versus time curve of the test drug (or route) to that of the standard when equal doses are administered:

$$F = \frac{AUC_{\text{test route}}}{AUC_{\text{iv}}}, \tag{9}$$

where AUC is the area under the plasma concentration versus time curve. If repeated doses are given, F can be calculated from the ratio of the steady-state concentrations C_{ss} which result from the same dose being given via the test route and i.v.:

$$F = \frac{C_{\text{ss,test}}}{C_{\text{ss,iv}}}. \tag{10}$$

Actual rates of transfer across biological barriers are generally calculated from Eqs. (5) and (6). To calculate the rate coefficients of either equation, data consisting of concentration and volume versus time in both the compartment in which the drug is administered and in the receiving compartment are required. If the target compartment is the circulation, venous concentrations may be the only data available. In this case, the rate of transfer can be expressed as the product of the plasma concentration and the volume of distribution divided by the total dose administered. This produces a fractional rate of absorption defined by

$$\text{fractional rate of absorption} = \frac{C_{\text{plasma}} V_{\text{d}}}{\text{dose} \times \text{time}}. \tag{11}$$

III. DRUG DISTRIBUTION

Once a drug enters the circulation of an organism, the drug is mixed with the fluid circulating through the body. After absorption of a drug into the bloodstream, it is simultaneously distributed throughout the body and eliminated. In mammals such as humans, the circulation consists of the arteries, veins, and the heart. The lymphatic circulation represents a third circulation, which returns proteins, salts, and fluid from the interstitium to the venous circulation; the flow rate in this system is slow relative to the blood and is therefore less important for low-molecular weight substances which rapidly transport across blood endothelia. However, the lymphatics can be significant in the recirculation of high-molecular weight solutes in the body; the daily flow through the thoracic duct amounts to 1–2 plasma volumes per day.

The distribution phase within the central volume of distribution V_{d} is usually assumed to be very short relative to the half-life of the drug $t_{1/2}$. Unbound, highly lipid-soluble medications cross cell boundaries rapidly and may be quickly metabolized by the liver or distributed to fat. Medications which are highly bound (>99%) may be restricted to the vascular space but may have a large apparent volume of distribution. Most water-soluble medications, either unbound or partially bound, are distributed rapidly via the circulation to all regions of the body and, depending on the nature of the endothelial barrier, transport to some degree into the extravascular space of tissue. The drug may transport from the extravascular space into a cell; it may be bound locally; or it may undergo recirculation to the vascular space, where it will undergo distribution again or be eliminated via several mechanisms discussed below. Data concerning binding, volume of distribution, and mechanism of elimination on many drugs are tabulated in useful, pocket-sized references which are listed in the Bibliography.

There may exist other compartments which exchange with the central circulation which are not well perfused but must be accounted for in modeling the kinetics of a drug (see Fig. 6). Muscle in a recumbent, resting human is relatively poorly perfused compared to the liver or kidney. General expansion of the muscle interstitium is termed *edema* and can significantly alter the magnitude of V_{d} in the case of intensive care unit patients who have received massive amounts of intravenous fluids. In certain other pathologic conditions, the extravascular space of patients may undergo extensive local increases and form so-called "*third spaces*." In the pleural space, this is called a *pleural effusion*; in the peritoneal cavity, the fluid is termed *ascites*; fluid around the heart is termed a *pericardial effusion*. Each of these fluid collections forms a compartment which can exchange with the normal extracellular

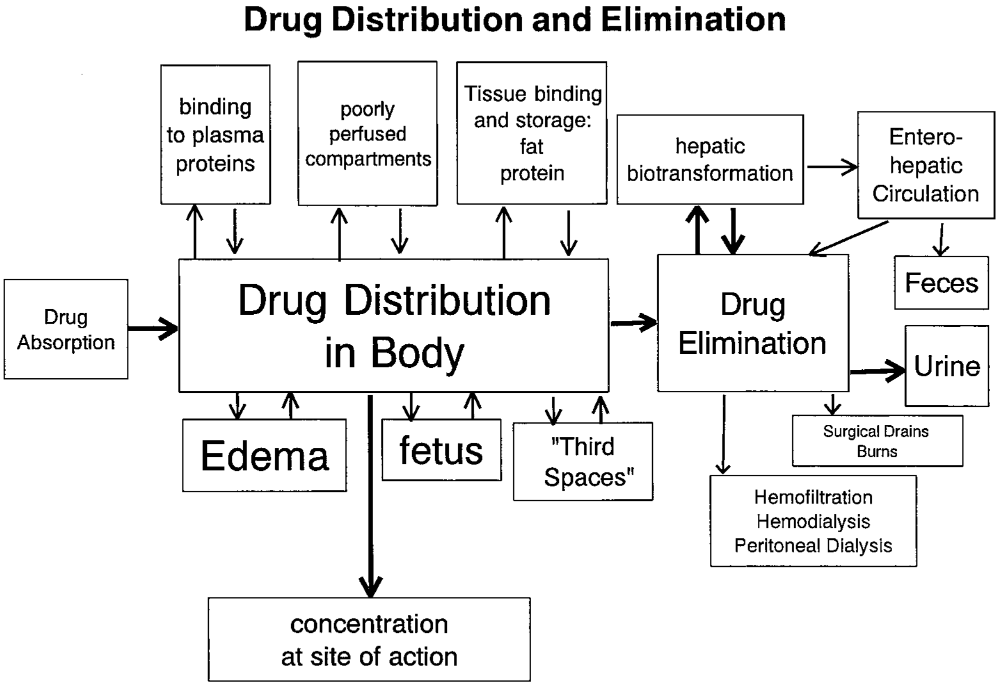


FIGURE 6 Drug distribution and elimination: detailed considerations. See text.

space and the circulation. Depending on the drug, these compartments may need to be included in detailed pharmacokinetic calculations for patients with these disease states.

A. Volume of Distribution

In order to calculate the dose needed to arrive at an effective drug concentration in the patient's plasma, the magnitude of V_d is needed ($C = \text{dose}/V_d$). To eliminate the need for collection of pharmacokinetic data on each patient, standard methods of calculating V_d with easily measured quantities in each patient have been developed. Since few patients weigh an ideal 70 kg, pharmacologists have attempted to correlate the loading dose with body weight or body surface area. Recent studies have found that lean body mass (LBM) is a better predictor of the body compartment size V_d than is the total body weight. LBM is defined as body cell mass, extracellular water, and nonfat intercellular connective tissue (bone, tendons, ligaments, basement membranes). LBM may be calculated as follows:

$$\text{LBM}_{\text{men}} = 1.10\text{TBM} - 120(\text{TBM/height})^2, \quad (12a)$$

$$\text{LBM}_{\text{women}} = 1.07\text{TBM} - 148(\text{TBM/height})^2, \quad (12b)$$

where height is in centimeters and LBM and total body weight or mass (TBM) is in kilograms. A related index is the ideal body weight (IBW), which is based on height:

$$\text{IBW}_{\text{men}} = 50\text{ kg} + 2.3\text{ kg} \times (\text{in. of height greater than 5 ft}) \quad (13a)$$

$$\text{IBW}_{\text{women}} = 45.5\text{ kg} + 2.3\text{ kg} \times (\text{in. of height greater than 5 ft}). \quad (13b)$$

Obesity is a major problem in over 50% of the adult population of the United States. Since part of the adipose tissue is water, significant amounts of fat can increase the V_d of even the most hydrophilic drug. For patients who are obese (>20% above ideal body weight), there is an increase in the effective LBM as well as in adipose tissue. To take this into account, 20–40% of the total body weight above the IBW should be added to the IBW to obtain the LBM. Surprisingly, moderately lipophilic drugs have not been shown to require adjustment in obesity, other than to take into account the additional LBM. However, some highly lipophilic drugs, such as phenytoin, verapamil, lidocaine, and the benzodiazepines, distribute to a significantly larger volume than LBM in cases of severe obesity, while others (cyclosporine, propanolol, and prednisolone) do not. The following equation is based on limited data and should be considered an approximate solution to the problem of accounting

for large amounts of fat in a body when calculating the LBM:

$$\text{LBM}_{\text{obesity}} = \text{IBW} + 0.3(\text{TBM} - \text{IBW}). \quad (14)$$

In some cases, the postsurgical or septic patient will become severely edematous and take on 10–20 L of fluid. Since this expansion occurs chiefly in the extracellular space, drugs will distribute to the edematous space. This additional weight must be added to the IBW, which can be estimated from the patient's height, or alternatively, can be added to the LBM calculated from the admission height and weight.

B. Binding

There is typically a large difference between the apparent volume of distribution, V_d, and the actual anatomic volume which contains the drug, V_{anat}. The volume V_{anat} cannot exceed the *total body water*, which is equal to approximately 60% of the lean body mass. The total body water is approximately two-thirds intracellular and one-third extracellular. The *extracellular space* is further divided into *intravascular* (7–8% of LBM, with the plasma taking up 4–5% and the red cells taking up the rest) and *extravascular space* (11–12% of the LBM). A drug with no binding will have V_d which matches one of these volumes. A drug which is highly bound to plasma proteins (>99%) may have an apparent volume of distribution many times its actual volume of distribution, which would approximate the intravascular space.

Binding to plasma proteins usually restricts movement of the drug out of the vasculature. Since *albumin* is the most important protein with respect to binding, states of altered serum albumin concentrations, such as pregnancy, liver dysfunction, or the nephrotic syndrome, can change the free plasma concentration of a highly bound drug and significantly influence the distribution and elimination of a drug. α_1-Acid glycoprotein is another important binding protein for basic drugs. Its molecular weight is 40,000 Da and it will therefore pass through the capillary endothelium. Since it is an acute-phase reactant, its concentration in plasma increases with states of inflammation or cancer but falls with pregnancy or nephrosis. Other proteins play a minor role in drug binding.

In addition to plasma proteins, cellular elements and tissues bind drugs. For example, it has been demonstrated that significant amounts of acetazolamide, phenytoin, propanolol, and quinidine are taken up in red blood cells. Drug binding to tissues is important but poorly understood, with the chief effect being on the time course of the drug. A decrease in tissue binding results in a decrease in the half-life of the drug.

Age, sex, and disease states such as hepatic cirrhosis, renal failure, and congestive heart failure affect drug binding and distribution. These will be discussed below.

IV. DRUG METABOLISM AND ELIMINATION

As illustrated in Fig. 2, once a drug is absorbed and distributed, it will enter the elimination phase. It may be removed directly by the kidney and appear unchanged in the urine. Or it can be taken up by the liver or another organ and metabolized to secondary products, which are secreted into the biliary system or are cleared by the kidney. If the patient is in renal failure, some form of renal replacement therapy will be initiated, and drug circulating in the plasma will be cleared to the dialysate according to the degree of plasma binding and the molecular size of the free drug.

A. Renal Excretion

Approximately 20% of the cardiac output circulates through the kidneys. As the blood passes through each glomerulus, a portion is filtered at the glomerular filtration rate (GFR). GFR decreases with age, and can be estimated from the Cockcroft–Gault relationship:

$$\text{GFR}_{\text{men}} = \frac{(140 - \text{age}) \times \text{IBW(kg)}}{72 \times \text{serum creatinine (mg/dl)}}. \quad (15)$$

The $\text{GFR}_{\text{women}}$ is found by multiplying Eq. (15) by 0.85. IBW is defined by Eqs. (13a) and (13b); age is in years; serum creatinine is measured. Solutes of low molecular weight (<6000 Da) which are not bound to protein are filtered at the same rate as plasma water. Larger molecules or those bound to proteins will pass through the glomerulus at a slower rate than the GFR. Macromolecules with anionic charges are further restricted in their passage through the normal glomerulus when compared with molecules with neutral or positive charges. The glomerular filtrate in the tubule is further modified in the proximal tubule by reabsorption or secretion of substances. Diuretic medications such as hydrochlorothiazide, furosemide, and amiloride are secreted in the proximal tubule and act on transporters located on the apical side of the tubule further along in the nephron.

For a drug whose transport is not limited by renal blood flow, the following expression can be used to calculate renal clearance (Clr_{kidney}):

$$Clr_{\text{kidney}} = f_d(\text{GFR} + Clr_{\text{KS}})(1 - f_R), \quad (16)$$

where f_d is the free drug fraction in blood, Clr_{KS} is the rate of renal tubular secretion of drug, and f_R is the fraction

of filtered or secreted drug which is reabsorbed in the tubules. The pH of the urine can affect the secretion, reabsorption, and overall excretion of ionized compounds; an acidic pH promotes excretion of weak bases, while an alkaline pH will promote excretion of weak bases. If there is no secretion or reabsorption, Eq. (16) simplifies to $Clr_{\text{kidney}} = f_{\text{d}} \times \text{GFR}$.

For substances whose transport is limited by the blood flow in the kidneys, Eq. (16) becomes

$$Clr_{\text{kidney}} = (f_{\text{d}} \times \text{GFR} + \text{RBF})(1 - f_{\text{R}}), \tag{17}$$

where RBF is the renal blood flow.

B. Hepatic Metabolism

Transformation mechanisms of drugs by the liver include metabolism, detoxification, and biotransformation. Most metabolism in the body occurs in the hepatic microsomes, a cellular fraction of the endoplasmic reticulum. The four major processes of metabolism are oxidation, reduction, hydrolysis, and conjugation. *Oxidation* is carried out by the microsomal mixed-function oxidases including cytochrome P450. *Reduction* involves the biotransformation of drugs with azo linkages (RN=NR′), nitro groups (RNO_2), and carbonyl compounds (RCOR′). Drugs which contain ester functions (RCOOR′), amides, or peptides are *hydrolyzed* in the liver. These three processes often produce a compound which is then *conjugated* with glucuronic acid or glutathione for excretion.

Often a single drug will have many metabolites including some which have effects similar to the parent (cyclosporin, chlorpromazine). Microsomal drug metabolism can be stimulated by medications such as the barbiturate phenobarbital or by cigarette smoke. Metabolism may be slowed by medications such as the monoamine oxidase inhibitors which are used in treatment of psychiatric disease. There are genetic and sex-related differences as well as age-related effects which may affect an individual patient's metabolism.

Michaelis–Menten kinetics is often used to describe an enzymatic process and can be described mathematically by Eq. (6) [rate $= V_{\max}C/(K_{\text{m}} + C)$]. In this case, C is the drug concentration in plasma, $V_{\max}$ indicates the total amount of metabolizing enzyme, K_{m} is the Michaelis constant , and $1/K_{\text{m}}$ is a measure of the affinity between the drug and the enzyme. If $C \ll K_{\text{m}}$, Eq. (6) reduces to:

$$\text{rate of metabolism} = \frac{V_{\max}}{K_{\text{m}}}C = k_{\text{m}}C, \tag{18}$$

where k_{m} is the apparent first-order metabolic rate constant, which can be used to calculate rates of metabolism.

The liver plays the major role in metabolism of substances in the blood stream. This is why the "first-pass effect" is so important to pharmacokinetics and the ultimate effect of a certain dose of drug. Most reactions which result in conjugation of the drug or metabolite decrease the pharmacologic activity of the drug. Thus, administration of a drug via a route without an initial pass through the liver will usually result in a greater effect for a given drug dose.

C. Extrahepatic Metabolism

Although the liver is the major organ of metabolism, other tissues, such as mucosa of the gastrointestinal tract, kidney, lung, brain, and skin also possess enzymes which metabolize drugs. Indirect evidence of extrahepatic metabolism is derived from studies in which observed rates of metabolism exceed the maximum rate of liver blood flow or in which severe liver disease or anhepatic conditions have not affected metabolic clearance. Enzymes located in the intestinal mucosa are likely the reason why orally dosed medications such as isoproterenol, terbutaline, or albuterol must be administered in much larger doses than if given i.v. The kidney carries out conjugation reactions (such as glucuronidation) and sulfation in the cortex and medulla. The lung has a high affinity for basic amines such as meperidine, lidocaine, and fentanyl. Brain contains monoamine oxidase as well as enzymes for ketone production; P450 enzyme levels are highest in the brain stem and cerebellum. The skin possesses several enzyme systems which hydrolyze, reduce, and convert steroids to active or inactive factors. All of these disparate sites play a minor role in relation to the normal liver. However, in severe liver dysfunction, their role in drug metabolism may become prominent.

D. Biliary Excretion

Drugs taken up by hepatocytes may be secreted into the bile and flow into the intestine. The drug or metabolite may either be excreted in the feces or be reabsorbed in the intestine in the process of enterohepatic recycling. This recycling may result in the persistence of the drug and active metabolites in the body. Some drugs are concentrated in the bile with bile to plasma drug concentration ratios of 1–1000. Since biliary flow is approximately 0.2–0.5 ml/min, the effective biliary clearance can be as high as 500 ml/min. If, for example, bile is surgically drained via a T-tube, clearance of a drug which is normally recycled could be considerable, and the half-life of the drug would be markedly decreased. Cardiac glycosides undergo extensive enterohepatic recycling, and the half-life of digitoxin can be significantly decreased by biliary drainage or administration of cholestyramine (which absorbs bile salts and prevents their recycling). Chemical

structure, polarity, and molecular weight are important factors in determining whether a drug is excreted in the bile.

V. PHARMACOKINETIC VARIABILITY

Age, sex, altered physiologic states such as pregnancy, and disease states such as hepatic cirrhosis, renal failure, and congestive heart failure can produce significant changes in drug binding, distribution, and elimination. Data are often limited in specific patient conditions, but the following summarizes selected topics.

A. Pharmacokinetic Changes in the Elderly

As the populations of Western countries age, the changes which occur with additional years of life have received more attention. Studies of healthy adults have demonstrated that there is a linear decrease in the following physiologic variables from age 30 to 90: nerve conduction velocity, basal metabolic rate, standard cell water, cardiac index, glomerular filtration rate [see Eq. (15)], renal plasma flow, vital capacity, and maximum breathing capacity. It is not surprising that there may be changes in drug absorption, distribution, metabolism, excretion in the urine, or biological response to the drug.

While there is evidence that aging likely results in reduced gastric motility, increased gastric pH, decreased surface area of small intestine, and reduced portal circulation, there is no evidence that these factors limit absorption after oral ingestion. However, drug distribution and elimination may change as a patient ages. Protein binding is quite variable in the elderly, with over-65-year-old patients demonstrating significantly less binding of warfarin and phenytoin than patients 20–40 years old. However, studies of diazepam, sulfadiazine, and phenylbutazone revealed no difference between the elderly and younger patients. With aging, the lean body mass tends to decrease, while the proportion of adipose tissue increases. Thus, an unbound, lipid-soluble drug may have a larger apparent volume of distribution in an elderly individual than in a younger person, while water-soluble drugs may be more concentrated with a smaller volume of distribution. Because of diminished GFR in the elderly [see Eq. (15)], the plasma half-life of an antibiotic may be prolonged and the urinary concentration diminished. If the antibiotic is prescribed to treat a urinary tract infection, urinary concentrations may be insufficient to completely treat the infection and this could cause recurrent or incompletely treated infections in the elderly.

Since cardiac output declines 1%/year between ages 19 and 86, regional blood flow to a variety of tissues may change, and the preferential distribution of flow to the brain, heart, and muscle may result in drug accumulation in these organs. Both autopsy studies as well as *in vivo* ultrasound examinations have shown that between 20 and 80 years of age, liver size decreases 18–24%. The decrease in liver size may be a major factor in the decreased elimination of drugs which are metabolized by enzyme systems with limitations in capacity. Clearance studies of indocyanine green have demonstrated that liver blood flow decreases 35% in those over 65 years and that the blood flow per unit of liver volume falls by 11%. However, due to lack of long term-data, the clinical significance of these changes in a normal elderly individual is uncertain.

B. Transplacental Transfer of Drugs

Due to ethical restraints, *in vivo* human data of drug transfer from mother to fetus are very sparse. Because of the lack of knowledge of toxicity to the fetus of many agents, drug usage must be restricted to those medications which are absolutely necessary to treat the mother and the child. Maternal drug absorption will be altered by the physiologic changes during pregnancy including decreased intestinal motility, a slowing of gastric emptying during late term, and emesis during early pregnancy. These may all contribute to a reduced absorption, a delay in peak concentration, and a reduced peak concentration. During pregnancy, the plasma volume increases with concomitant changes in red cell volume, and total body fat usually increases. Although binding affinities remain the same, the protein concentration decreases. These alterations affect the volume of distribution of drugs. Metabolism is affected by alteration of the proportion of cardiac output which flows to the liver. Renal elimination during pregnancy increases for drugs which are cleared by glomerular filtration, but secretion by renal tubules is more variable.

Transfer across the placenta is dependent on a number of factors: (a) surface area of the membrane, (b) thickness of the membrane, which thins from 50–100 μm at 2 months to 4–5 μm at term, (c) maternal blood flow and intervillous blood pressure, (d) blood pressure in fetal capillaries, and (e) the fetal–maternal osmotic gradient. A large number of substances such as pituitary and thyroid-stimulating hormones, insulin, corticotrophin, amines, and low-molecular weight heparins are incapable of crossing the placental membrane. Unbound, nonionized low-molecular weight substances transfer by passive diffusion, while large antibodies cross by endocytosis. Higher lipid solubility also favors transfer. The standard reference substance in this class is phenazone (antipyrine), and the clearance of a given substance is often listed as a clearance index or the ratio of the fetal transfer rate of drug “x” to that of phenazone.

Once across the placenta, a drug distributes to the fetus. During growth to term, fetal body water decreases while fat volume increases from near zero to 12% at term. The amount of protein for binding of drugs and their affinity are reduced in the fetus. Metabolism of some drugs by the fetal liver is almost nil, while for others it may increase to over 30% of the adult capacity. Although fetal urine flow increases progressively through gestation, because of oral recirculation, most of the elimination is through the maternal circulation. Unfortunately, data are available on only a limited number of drugs. Extrapolations from substances with similar physical properties may produce large errors of estimation.

C. Renal Failure

Renal insufficiency, whether due to pre-, post-, or intrinsic renal causes, will result in the decreased elimination of many drugs. Often, the physician may anticipate a decreased renal function from unanticipated high trough levels of a drug such as the antibiotic gentamicin, even if the increased drug levels occur prior to an obvious rise in serum creatinine. For a substance primarily excreted unchanged via the kidney, intrinsic renal disease such as in acute tubular necrosis would cause a decrease in GFR and perhaps a marked decrease in secretion (Clr_{KS}). From Eq. (16), this would cause a proportionate decrease in clearance and result in a longer half-life. If the drug is blood flow-limited, severe hypotension would result in a decreased RBF and a decrease in clearance, in accordance with Eq. (17). In some cases, renal failure does not result in the decreased clearance of a highly bound substance which is primarily excreted by the kidney. As the fraction of unbound drug increases due to renal insufficiency, the increase in f_d offsets the decrease in GFR and Clr_{KS} in Eq. (16).

Uremia has a variable effect on hepatic metabolism. In uremic rats, demethylation enzymes and cytochrome P450 decrease their activity 30–40% compared to levels in nonuremic rats. On the other hand, alcohol dehydrogenase activity is increased in uremia 50–90% above normal physiologic conditions, but uremia produces inhibiting factors of hepatic systems. Indeed, the nonrenal clearances of many drugs used daily in hospitals such as the cephalosporins, imipenem, metoclopramide, or procainamide are significantly decreased by end-stage renal disease (ESRD). The half-life of many anxiolytic drugs such as benzodiazepines are markedly prolonged in renal failure. The dosage interval must be adjusted carefully. Guidelines for dosing many commonly used medications are typically based on ranges of GFR or creatinine clearance and are readily available in pocket-size form (e.g., Bennett *et al.*, 1994; Chernow, 1995).

D. Removal of Drugs by Dialysis

Patients with less than 10% of normal kidney function require renal replacement therapy for removal of waste metabolites. In patients undergoing hemodialysis, the total clearance of a drug is equal to sum of the clearances due to nonrenal routes of elimination (Clr_{NR}), residual renal function (Clr_{RRF}), and the dialyzer ($Clr_{dialyzer}$):

$$Clr_{total} = Clr_{NR} + Clr_{RRF} + Clr_{dialyzer}. \qquad (19)$$

Intermittent hemodialysis is the most widely used modality of renal replacement therapy in the United States. Unbound low-molecular weight drugs are cleared from the blood in the same way as creatinine. The absolute mass rate of drug removal can be estimated by measuring the drug concentration at the inlet (C_{inlet}) and outlet (C_{outlet}) of the dialyzer and multiplying the difference by the blood flow rate (Q_B) through the dialyzer (rate $= Q_B \times (C_{outlet} - C_{inlet})$). The clearance will depend on the type and size of the artificial membrane, properties of the drug molecule, and blood and dialyzate flow rates. If measurement of drug levels are not an option, the $Clr_{dialyzer}$ can be estimated from characteristics of the membrane for substances of similar molecular weight by the following equation:

$$Clr_{dialyzer} = Clr_{est} \times f_d \left[1 - \left(\frac{HCT}{100}\right)\right], \qquad (20)$$

where Clr_{est} is the estimated dialyzer clearance from dialyzer manufacturer specifications, HCT is the hematocrit, and f_d is the unbound fraction of drug. If the physician does not have the necessary information to use Eq. (20), then levels must be closely monitored in the case of drugs which are toxic at high concentration. Drugs with minimal toxicity must be dosed with the assumption that in the "standard" dialysis treatment a certain percentage of drug is cleared. This latter method is the basis of tabulations of recommended dosing after dialysis.

Drugs are cleared from the blood during peritoneal dialysis in an analogous fashion to the clearance of creatinine or other plasma constituents. Transport between the blood and the peritoneal cavity is symmetric in the sense that transport rates are equivalent in each direction, and therefore if the PA (mass transfer–area coefficient) for a substance of equivalent molecular weight is available, the transfer rate can be estimated from

$$\text{drug transfer rate} = PA \times (C_B - C_{PC}). \qquad (21)$$

Clearances and recommendations on dosing are tabulated in several references. All of these tabulations assume that the peritoneum is normal; inflammation which results from acute peritonitis can cause significant increases in rates of transport.

E. Patients with Liver Disease

Most data concerning drug elimination in liver disease have been obtained in patients with cirrhosis. Current theories regard the decreased permeability of the hepatic sinusoid as a major feature in the changes brought about in the liver. Oxidative metabolism is typically impaired in cirrhosis with the sparing of drug glucuronidation until severe impairment occurs. Biliary excretion is also impaired. Moderate liver impairment decreases the clearance of the native drug or metabolites from the kidney, often in spite of normal creatinine clearance. Because renal tubular secretion of creatinine is increased in cirrhosis, creatinine clearance may not reflect the actual degree of renal dysfunction. In cases of liver disease without cirrhosis (carcinoma, viral hepatitis), dose changes are probably not necessary. In general, chronic liver disease with cirrhosis requires reduction in dose, regardless of the route of elimination.

F. Congestive Heart Failure

Congestive heart failure includes cardiac "pump" failure and results in poor perfusion of the liver and kidney. Congestion also occurs in the liver and the gut. Pharmacokinetics is altered in heart failure: the volume of distribution is reduced and clearance is decreased. The net result of changes is often not predictable, but plasma drug concentrations are usually higher in patients with congestive failure than in healthy persons. Drug levels of such common cardiac drugs as lidocaine rise proportionately to the degree of cardiac failure. Plasma concentrations in cardiac shock can rise to toxic levels. The half-life of antipyrine, used as a marker of hepatic metabolism, is prolonged in cardiac failure but returns to normal during convalescence. It is therefore important to monitor drug levels of toxic drugs during conditions of myocardial infarction and severe congestive failure.

G. Patients with Burns

Burn injuries cause a variety of changes in drug pharmacokinetics. Drug absorption of orally administered drugs may be enhanced because of increased permeability of the intestine. Unfortunately this condition permits bacteria to pass into the blood as well and often results in septicemia or shock. For the first few days, the levels of proteins do not change, but after 4–5 days, albumin progressively falls and α_1-acid glycoprotein (AAG) increases. Decreased albumin concentrations but increased AAG levels affect the distribution and elimination of drugs. The binding has major effects on the clearance and V_d. Clearance is often enhanced directly through the burned area. In addition, renal clearance appears to be increased 1 week after the injury, while hepatic metabolism is quite variable. Poorly extracted drugs may have a significantly decreased hepatic clearance. Therapeutic drug monitoring is recommended for all critical medications in burn patients.

VI. QUANTITATIVE APPROACHES TO PHARMACOKINETIC MODELING

The goal of pharmacokinetics is the quantitative description of drug entry, distribution, and elimination in the body. These processes are typically integrated into a mathematical model which uses a system of equations such as Eqs. (1)–(20) to calculate the systemic concentration due to a certain dose administered to the patient. If the pharmacodynamic characteristics of the drug can be clearly defined, the desired concentration at the target or in the plasma can be specified, and the pharmacokinetic model can be used to calculate the dose to attain the effective concentration.

A. Compartmental Approaches

The classic approach to pharmacokinetic modeling is to describe the system as a group of arbitrarily sized compartments with one compartment designated as the central compartment which receives the drug via some route and from which the excretion occurs. Figure 7 displays a three-compartment model. In this example, we will assume that the drug has been given directly into the central compartment (1); however, a fourth compartment could be added for the route of absorption or an input function of drug mass entering compartment 1 versus time. The compartments may or may not match anatomic compartments such as the plasma space or the extracellular space. The peripheral compartments 2 and 3 may represent parts of the body which have different rates of perfusion than the central compartment. Each compartment is considered

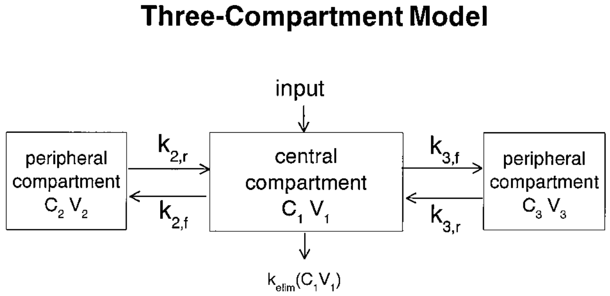


FIGURE 7 Three-compartment model. The k_i are the rate coefficients, C_i is the concentration in compartment, i, and V_i is the volume in compartment i. See text for detailed discussion.

to be well mixed with a uniform concentration throughout. The concentrations $C_i(t)$ and volumes $V_i(t)$ can all be functions of time t. The mass balance around the central compartment is

$$\frac{d(C_1V_1)}{dt} = R_{\text{inf}} - (k_{3,\text{f}} + k_{\text{elim}} + k_{2,\text{f}})(C_1V_1) + k_{3,\text{r}}(C_3V_3) + k_{2,\text{r}}(C_2V_2), \qquad (22)$$

where R_{inf} is the rate of infusion or rate of delivery, which can be constant or dependent on time. The k's are rate coefficients which must be determined from curve fits to experimental data.

The mass balances for the other two compartments are as follows:

$$\frac{d(C_2V_2)}{dt} = k_{2,\text{f}}(C_1V_1) - k_{2,\text{r}}(C_2V_2), \qquad (23)$$

$$\frac{d(C_3V_3)}{dt} = k_{3,\text{f}}(C_1V_1) - k_{3,\text{r}}(C_3V_3). \qquad (24)$$

Often, only the infusion rate, the elimination rate ($k_{\text{elim}}C_1V_1$), and the volume and concentration of the central compartment versus time are known. This leaves eight unknowns with only three equations. The number of variables can be reduced to four by redefining mass ($M_i = C_iV_i$) and equating $k_{i,\text{f}}$ to $k_{i,\text{r}}$. Numerical fits to the data can produce a range of possible parameters in order to fit the central-compartment concentration data.

The compartmental model is primarily a mathematical scheme to predict the plasma or central compartment concentration. It does not tell us about the mechanisms inside the body which control drug distribution and elimination.

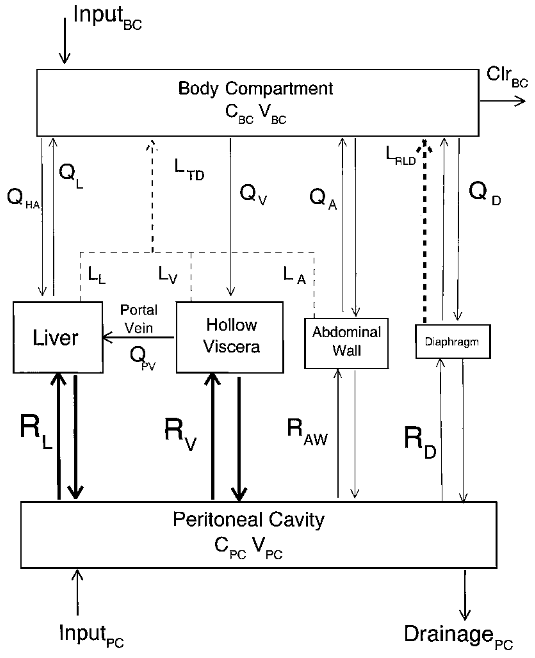


FIGURE 8 Physiologic model for transport between the blood (body compartment) and a solution in the peritoneal cavity. The model emphasizes the importance of the tissue surrounding the cavity. Each tissue compartment can be characterized by the mass of drug or the drug concentration and tissue-specific volume of distribution. Input of the drug can be intravenous or intraperitoneal. (i.p.). Drugs administered i.p. are typically done in the setting of peritoneal dialysis in which a portion of the drug is typically not absorbed but drained out of the cavity. Drugs are cleared from the plasma (body compartment) at a clearance rate Clr_{BC}. Here Q_i is the blood flow to or from organ i, L_i is the lymph flow rate, and R_i is the rate of mass transfer between peritoneal cavity and the tissue in contact with the peritoneal solution.

B. Physiologic Pharmacokinetic Models

In this modeling approach, anatomically or physiologically defined spaces within the body, along with blood flows to and from each, rates of drug extraction, or metabolism by each compartment are modeled. These models require significantly more detailed information about a system than the typical compartmental approach. An example of such an approach is illustrated in Fig. 8, a multicompartmental, physiologic model of exchange between the body and fluid in the peritoneal cavity. The transport of substances between the body compartment (the volume of drug distribution within the body with which the plasma is in equilibrium) and the peritoneal cavity occurs via the tissue compartments which surround the peritoneal cavity. Drugs may be introduced into the peritoneal cavity (Input$_{\text{PC}}$) or into the body compartment (Input$_{\text{BC}}$). From the body compartment, they can distribute to the tissue compartments around the cavity via the blood flow to each tissue group (Q_i). From the peritoneal cavity, the drug transports into the tissue compartments and then is taken up by the blood circulation (Q_i) or lymph circulation (L_i) and transports to the body compartment. That the hollow viscera drain directly into the liver via the portal vein is included in the model; if the drug is metabolized in the liver, this can be included in a submodel of the liver. Each flow rate, compartment volume of distribution, and rate of drug metabolism must be specified in such a model. Mass balances are written for each compartment and are solved simultaneously to estimate concentrations in each compartment. Since it has recently been shown that the extracellular volume of the tissues surrounding the peritoneal cavity expand when large volumes of fluid are infused into the cavity, volume balances must be written and solved to calculate the volume of each compartment.

Physiologic models are very complex and require detailed data to implement. The complexity of such models, however, provides the capability of studying the effects of variations in parts of the transport system. For example, the role of lymphatic transport of solute from the peritoneal cavity to the body compartment could be investigated by

setting all L_i to zero and recalculating the tissue concentrations and volumes. Thus this type of model is most valuable in the investigation of detailed mechanisms of absorption, distribution, and elimination in the pharmacokinetic simulation of a drug.

C. Noncompartmental Model Approaches

Noncompartmental pharmacokinetics has been developed as an alternative to data-intensive compartmental and physiologic models. While the latter techniques are useful in pharmacokinetic predictions if sufficient data are available, drugs with complex distribution and elimination may be difficult to properly model without additional experimental data. The noncompartmental techniques do not rely on specific distribution characteristics of a drug and therefore become useful when data are limited.

The basis for noncompartmental methods for calculation of the parameters of each step of absorption, distribution, and elimination is the *theory of statistical moments.* The information required is the drug concentration in the central compartment versus time with concentrations taken past the absorptive phase and distributive phase of the curve. The area under the concentration versus time curve (AUC) is the zero moment. The first moment of the AUC is the area under the curve of the product of the concentration times time versus time (M^1_{AUC}):

$$M^1_{AUC} = \int_0^\infty (Ct)\, dt. \tag{25}$$

The mean residence time ($\overline{t_{\text{residence}}}$) provides some estimate of how long the drug may last in the central compartment and is analogous to the drug half-life $t_{1/2}$. It is calculated as follows:

$$\overline{t_{\text{residence}}} = \frac{M^1_{AUC}}{AUC}. \tag{26}$$

The clearance of the drug from the central compartment can be calculated by

$$Clr = \frac{\text{dose}}{AUC}. \tag{27}$$

If a drug is infused at a constant rate of I_{constant} and the concentration stabilizes at a steady state C_{ss}, then the clearance can also be calculated by

$$Clr = \frac{I_{\text{constant}}}{C_{\text{ss}}}. \tag{28}$$

The apparent steady-state volume of distribution can be calculated from

$$V_{\text{ss}} = \frac{\text{iv dose} \times M^1_{AUC}}{AUC^2}. \tag{29}$$

The rate of absorption can be characterized from the calculation of $\overline{t_{\text{residence}}}$ for both i.v. administration and non-i.v. administration. The mean absorption time ($\overline{t_{\text{absorption}}}$) is calculated as the difference of the two mean residence times.

This method is chiefly descriptive and requires no understanding of the underlying mechanisms. It permits quantitative characterization of the kinetics of the drug in the central compartment. The advantage is that sophisticated mathematics is unnecessary. This fact alone makes noncompartmental methods particularly useful in the clinical use of drugs.

SEE ALSO THE FOLLOWING ARTICLES

ABSORPTION • BIOENERGETICS • ION TRANSPORT ACROSS BIOLOGICAL MEMBRANES • PHARMACEUTICALS, CONTROLLED RELEASE OF

BIBLIOGRAPHY

Aarons, L. (1999). “Software for population pharmacokinetics and pharmacodynamics,” *Clin. Pharmacokinet.* **36,** 255–264.

Bennett, W. M., Aronoff, G. R., Golper, T. A., Morrison, G., Brater, D. C., and Singer, I. (1994). “Drug Prescribing in Renal Failure,” 4th ed., American College of Physicians, Philadelphia.

Berner, B., and John, V. A. (1994). “Pharmacokinetic characteristics of transdermal delivery systems,” *Clin. Pharmacokinet.* **26,** 121–134.

Bourget, P., Roulot, C., and Fernandez, H. (1995). “Models for placental transfer studies of drugs,” *Clin. Pharmacokinet.* **28,** 161–180.

Bourne, D. W. A. (1997). “Using the Internet as a pharmacokinetic resource,” *Clin. Pharmacokinet.* **33,** 153–160.

Bressolle, F., Kinowski, J.-M., Emmanuel de la Coussaye, J., Wynn, N., Eledjam, J.-J., and Galtier, M. (1994). “Clinical pharmacokinetics during continuous haemofiltration,” *Clin. Pharmacokinet.* **26,** 457–471.

Chernow, B. (1995). “Critical Care Pharmacology,” Williams and Wilkins, Baltimore, MD.

Crooks, J., O’Malley, K., and Stevenson, I. H. (1976). “Pharmacokinetics in the elderly,” *Clin. Pharmacokinet.* **1,** 280–296.

Flessner, M. F., and Dedrick, R. (1999). “Intraperitoneal chemotherapy,” *In* “The Textbook of Peritoneal Dialysis” (R. Gokal, K. D. Nolph, and R. Krediet, eds.), pp. 769–790, Kluwer, Dordrecht.

Gex-Fabry, M., and Balant, L. P. (1994). “Considerations on data analysis using computer methods and currently available software for personal computers,” *In* “Pharmacokinetics of Drugs” (P. G. Welling and L. P. Balant, eds.), pp. 507–527, Springer-Verlag, Berlin.

Gibaldi, M. (1991). “Biopharmaceutics and Clinical Pharmacokinetics,” 4th ed., Lea and Febiger, New York.

Gibaldi, M., and Perrier, D. (1982). “Pharmacokinetics,” 2nd ed., Marcel Dekker, New York.

Gibaldi, M., and Prescott, L. (eds.). (1983). “Handbook of Clinical Pharmacokinetics,” Adis Health Science Press, Balgowlah, Australia.

Gibson, T. P. (1986). “Renal disease and drug metabolism: An overview,” *Am. J. Kidney Dis.* **8,** 7–17.

Guy, R. H., and Maibach, H. L. (1983). “Drug delivery to local subcutaneous structures following topical administration,” *J. Pharm. Sci.* **72,** 1375.

Hebbard, G. S., Sun, W. M., Bochner, F., and Horowitz, M. (1995). "Pharmacokinetic considerations in gastrointestinal motor disorders," *Clin. Pharmacokinet.* **28,** 41–66.

Holford, N. H. G. (1996). "A standard size for pharmacokinetics," *Clin. Pharmacokinet.* **30,** 329–332.

James, W. P. T. (1976). "Research on Obesity," Her Majesty's Stationery Office, London.

Kalant, H., and Roschlau, W. (eds.). (1998). "Principles of Medical Pharmacology," 6th ed., Oxford University Press, Oxford.

Leving, R. R. (2000). "Pharmacology: Drug Actions and Reactions," 6th ed., Parthenon, New York.

Morgan, D. J., and Bray, K. M. (1994). "Lean body mass as a predictor of drug dosage. Implications for drug therapy," *Clin. Pharmacokinet.* **26,** 292–307.

Morgan, D. J., and McLean, A. J. (1995). "Clinical pharmacokinetic and pharmacodynamic considerations in patients with liver disease," *Clin. Pharmacokinet.* **29,** 370–391.

Paton, T. W., Cornish, W. R., Manuel, M. A., and Hardy, B. G. (1985). "Drug therapy in patients undergoing peritoneal dialysis: Clinical pharmacokinetic considerations," *Clin. Pharmacokinet.* **10,** 404–426.

Peck, C. (1985). "Bedside Clinical Pharmacokinetics," Pharmacometrics Press, Rockville, MD.

Reetze-Bonorden, P., Bohler, J., and Keller, E. (1993). "Drug dosage in patients during continuous renal replacement therapy," *Clin. Pharmacokinet.* **24,** 362–379.

Schwartz, M. L. (1974). "Antiarrhythmic effectiveness of intramuscular lidocaine: Influence of different injection sites," *Clin. Pharmacol. Ther.* **14,** 77.

Shammas, F. V., and Dickstein, K. (1988). "Clinical pharmacokinetics in heart failure," *Clin. Pharmacokinet.* **15,** 94–113.

Wagner, J. G. (1971). "Biopharmaceutics and Relevant Pharmacokinetics," Drug Intelligence Publications, Hamilton, IL.

Ward, R. M. (1995). "Pharmacological treatment of the fetus," *Clin. Pharmacokinet.* **28,** 342–350.

Woodhouse, K. W., and Wynne, H. A. (1988). "Age-related changes in liver size and hepatic blood flow: The influence of drug metabolism in the elderly," *Clin. Pharmacokinet.* **15,** 287–294.

Separation and Purification of Biochemicals

Laure G. Berruex
Ruth Freitag
Swiss Federal Institute of Technology

I. Principles of Chromatographic Separations
II. Process Design in Chromatography
III. Stationary Phases for Biochromatography

GLOSSARY

Adsorption Interaction between biomolecules in the mobile phase and the surface of the chromatographic stationary phase, known as binding step. If the interactive phase is bonded to an inner core matrix—for example, the C_{18} phases used in reversed-phase chromatography—the distribution is called partitioning instead.

Batch operation A set volume of sample (=batch) is purified at a given time in a given column. Batch procedures are repeated in cycles, defining a certain time, cost, and productivity per cycle.

Biochromatography Liquid chromatography applied to biopolymers.

Biopolymer Biological macromolecule consisting of repeated "units" of biological origin, such as amino acids or nucleic acids. It does not mean a biocompatible polymer, or a biodegradable polymer.

CIP/SIP Cleaning and sanitizing operations carried out in place, part of the Quality Assurance procedures. CIP/SIP should be reproducible, validated, automated procedures done without opening or taking apart the concerned equipment.

GMP Good manufacturing practice is part of the quality system regulation to assure product consistency and quality. GMP includes requirements related to the methods used in, and the facilities and controls used for, designing, manufacturing, packaging, labeling, storing, installing, and servicing of medical devices and products intended for human use.

Isocratic elution Purification step consecutive to adsorption, consisting of collecting the target molecule (eluting) off the stationary at a constant mobile phase composition and ionic strength.

Isotherm The adsorption isotherm is the ratio of the times spent by a molecule in the mobile and the stationary phases, equal to the ratio of its concentrations in these phases. Usually defined for liquid/solid or gas/solid interactions.

Recombinant protein Protein obtained by recombinant technology (or genetic engineering), expressing a given

protein coded by its gene, in a host cell culture. The recombinant protein is identical to the one produced in the native source.

SOP Standard operating procedures are an important part of Quality Assurance and GMP in order to assure product quality and consistency.

BIOLOGICALS and especially the products of the modern biopharmaceutical industry come in many forms and from many sources. They can be molecules of various sizes ranging from small organic metabolites (ethanol, citric acid, antibiotics) to peptides and proteins (insulin, antibodies, enzymes) that are amino acid based, as well as most recently nucleotide-based molecules (antisense DNA, RNA, plasmid DNA). They may be metabolites of microorganisms, plant, or animal cells; they may be the product of an enzymatic reaction or a chemical peptide synthesis; or they may simply be found in a natural source such as blood, milk, or plant material. If, for example, the source is a microorganism, the desired substance may be excreted in the culture medium, may be enriched in the cyto- or periplasma, or may form inclusion bodies (large aggregates of the denatured protein found inside the bacterium). Product concentrations may range accordingly from several hundred grams per liter as in the case of ethanol or citric acid, down to a few milligrams per liter in the case of recombinant antibodies, and even micrograms per liter as in the case of blood coagulation factors. Given the present trend (mainly driven by economical considerations) the percentage of biopharmaceutical products such as recombinant proteins and antibodies for therapeutic use is expected to increase steadily over the next years. While the isolation/purification of a highly concentrated, low value product such as ethanol is relatively straightforward and based on standard operations in chemical engineering, the isolation of biologicals such as blood factors, monoclonal antibodies and recombinant proteins is much more demanding. The set of complex isolation and purification steps involved in the recovery of the product is generally referred to as the downstream process. In general, the contribution of downstream processing to the overall production cost depends on the upstream product concentration. It may vary from less than 10% in the case of citric acid to more than 90% in the case of some of the above mentioned high value products. As a result, a direct relationship has been observed between the selling price, respectively the production costs, and the product concentration in the original feed, as illustrated by Fig. 1.

In the design of the process leading to product recovery, the nature of the starting material is a key parameter. Equally important, however, is the desired quality of the final product, i.e., the maximum acceptable level and the chemical nature of impurities or contaminants. The requirements regarding purity for biologicals depend highly on the intended use of the product; they will be highest for human therapeutics and preventive drugs, somewhat less for a number of diagnostics, and often considerably lower for industrial enzymes. In the case of therapeutic proteins, requirements also depend on the administration dose, frequency, and route. Injectables have the highest requirements. Example values are given in Table I for a preventive injectable protein.

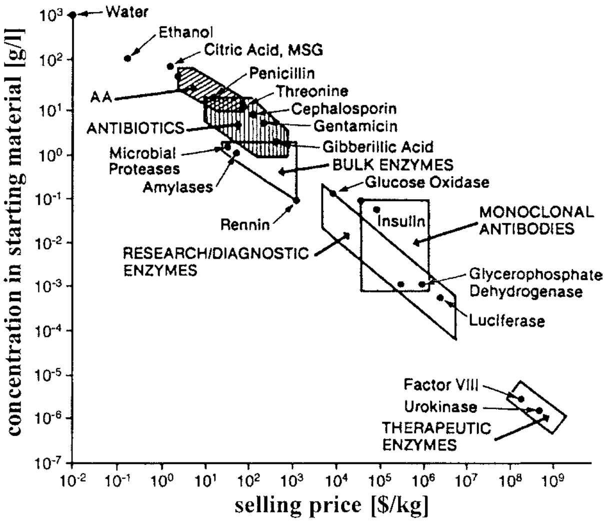


FIGURE 1 Relationship between the product concentration in the starting material and the selling price. [From Dwyer, J. L. (1984). *Biol. Technol.* With permission.]

Regulatory issues and environmental considerations (GMP, various national laws, etc.) are also important, especially in the case of industrial downstream processing. The

TABLE I Example of Regulatory Requirements for a Therapeutic Recombinant Protein[a]

Purity	>99%
Aggregates	<1%
Host cell proteins (HCP)	<1 ppm
Nucleic acids (mostly DNA)	<100 pg per dose
Endotoxins	<5 EU[b] per kg of body weight and per maximal dose per kg.
Viruses	>12 log reduction
Cells and particles	Not detectable
Leachables	<1 ppm

[a] Final purity requirements depend highly on the intended use of the product, diagnostic, preventive or therapeutic. No definite values can be given for therapeutic biomolecule, as they also vary depending on the administration frequency and route.

[b] EU = Endotoxin units (see FDA, 12/1987).

design and validation of standard operating procedures (SOP), between batch in-place cleaning (CIP), and sanitizing (SIP) procedures or the validation of operational parameters such as column leaching and the necessity to remove certain impurities (endotoxins, virus DNA, etc.) have to be taken into account. The main regulatory concern is the possibility of contamination of the product with any substance whose administration could be detrimental to the patient.

Figure 2 places the recovery process within a typical biotechnological production process. A typical recovery process can be roughly subdivided into four steps. First, the bioproduct is separated from the producing organisms and other insolubles by a solid/liquid separation step, such as centrifugation or filtration. This may require cell rupture in the case of intracellularly enriched substances or resolubilization in the case of inclusion bodies. In the second isolation step, substances that differ considerably in their physicochemical character are removed from the product. The methods used are either highly specific, based on biospecific (affinity) interactions or substance class specific such as salting out, solvent extraction, or batch adsorption procedures. If well designed, this step should result in a considerable increase in product concentration

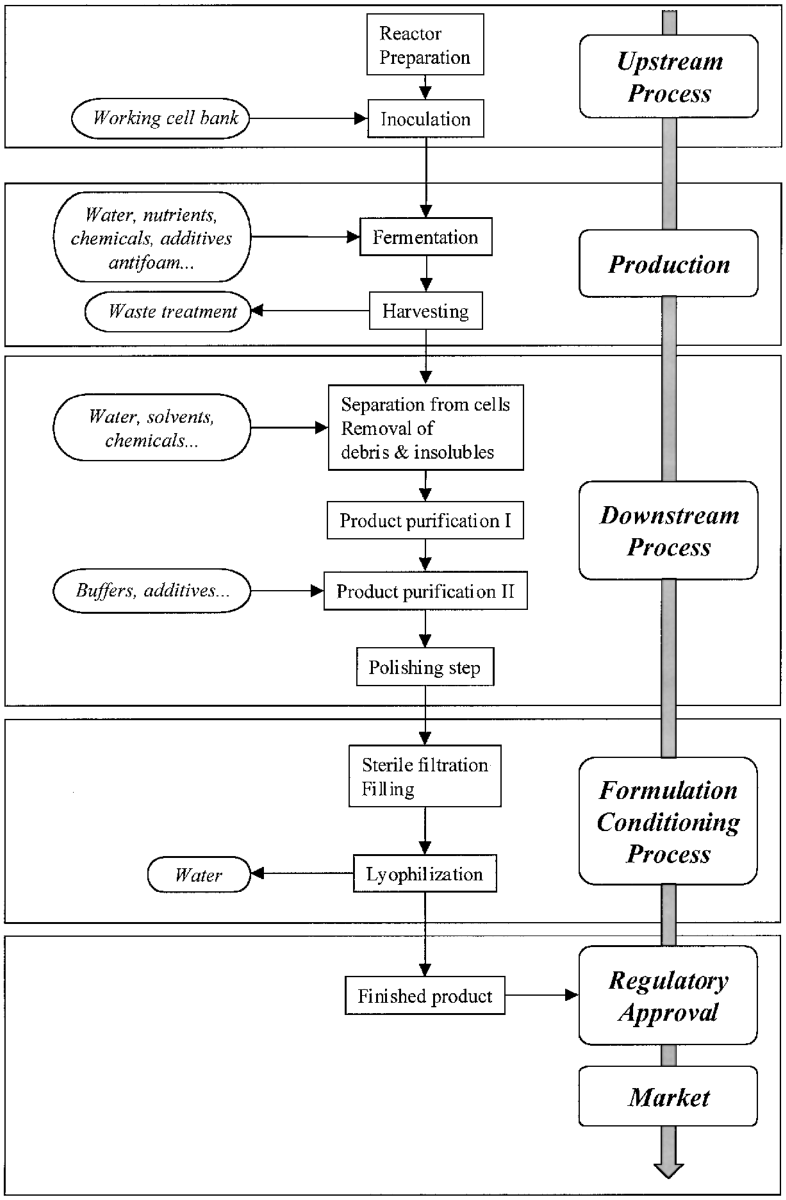


FIGURE 2 Downstream processing is an important part of a typical bioproduction process, involving preparative biochromatography at several stages.

to facilitate the following purification steps. The next steps often involve highly selective methods to remove substances of similar physical properties and biochemical functions from the product. At this point, chromatography plays the major role. The fourth step is polishing, and may include gel filtration, crystallization, and lyophilization of the final product. Chromatography has an important place in the methodologies used for downstream processing at nearly all levels. Electrophoretic methods have also been used in the past for the isolation of such target molecules, but they have been outmoded, replaced by the modern, faster chromatographic methods. Electrophoresis is still used nowadays, but mainly on the analytical scale in quality control. Biospecific affinity beads or magnetic beads, and affinity precipitation, are timidly entering the field on a preparative scale, for scavenging product in culture supernatant or sometimes directly in the culture media, by one-step adsorption processes, and will not be detailed here.

I. PRINCIPLES OF CHROMATOGRAPHIC SEPARATIONS

Chromatography has been the primary preparative separation method in biology and biochemistry since 1906, when the Italo-Russian botanist Tswett separated the pigments of chlorophyll by passing petroleum ether extracts of plant material through a column packed with powdered chalk. In the 1930s, column chromatography became an important tool for the separation of natural products and later, with the introduction of paper chromatography, for the microanalysis of biological samples. Various characteristics of biomolecules ranging from their general physicochemical properties (size, charge, hydrophobicity) to biospecific interactions have been exploited for their chromatographic separation. Although several modes of operation have been recognized in chromatography, most effort has been placed on the development of (linear) elution (*syn.*: desorption) chromatography. The development of high performance liquid chromatography (HPLC) in the 1970s allowed high-speed analysis, and set new standards of precision and resolution in the liquid chromatography of small molecules and beginning in the 1980s also of larger biologicals (proteins, DNA).

The following relationships are fundamental to all types of chromatography. A certain number of basic parameters allow the description and evaluation of a chromatographic separation, e.g., the calculation of the resolution from parameters such as the retention time and the zone widths. A complete chromatographic process can be described by the mass balance of the system, also allowing modeling of such processes. In addition, the number and kind of separation principles used in size-exclusion and interactive chromatography is limited and common to all modes of biochromatography. An overview of the different chromatographic modes and the set-up conditions such as column and mobile phase type is given in Table II. Table III compares the modes in terms of their suitability for a given application mode and some important factors to be considered during optimization.

TABLE II Main Characteristics of the Different Chromatographic Modes and Their Applications

Chromatographic mode	Stationary phase	Mobile phase	Applications
Size exclusion (SEC) or Gel filtration (GF)	Particles of well-defined size, of different sized pores	Aqueous buffer	Desalting, buffer exchange Determination of molecular weight Final polishing
Ion exchange (IEC)	Particles coated with anion or cation exchanger functionalities	Aqueous buffers, containing salts for elution	Separation of charged molecules Good choice for protein separation on preparative scale
Affinity (AC)	Covalently immobilized affinity ligand on particles	Aqueous buffers	Good choice for capture of target molecule at low concentrations in sample
Hydrophobic interaction (HIC)	Particles of well defined size, coated with small hydrophobic ligands C_2—C_4	Saline aqueous solutions	Well developed for protein separation No prior desalting necessary, good after IEC
Reversed phase (RPC)	Particles of well-defined size, coated with hydrophobic ligands C_4—C_{18}	Water, buffers of low molarity, and organic solvents for elution Presence of salts problematic above 10 mM	Indicated for neutral and uncharged molecules, soluble in aqueous/organic mixtures Excellent for analytical HPLC, seldom preparative (proteins denatured by organic solvents)

TABLE III Suitability of Purification Modes and Important Factors to Consider for Optimization

LC mode	Molecular characteristic	Main features	Purification steps	Sample start condition	Sample end condition	Important factors
SEC	Size	Limited resolution Low capacity Low speed	Intermediate (+) Polishing (+++)	Limited sample volume (<5% c.v.) and flow rate range	Buffer exchanged (if required) Diluted sample	Pore size and volume Bed height Flow rate
IEC	Charge	High resolution High capacity High speed	Capture (+++) Intermediate (+++) Polishing (+++)	Low ionic strength No volume limitation	High ionic strength or pH change Concentrated sample	pH Gradient slope Sample load
AC	Biospecific sites	High resolution Medium capacity High speed	Capture (+++) Intermediate (+++) Polishing (++)	Specific binding conditions No volume limitation	Specific eluting conditions Concentrated sample	Immobilization of ligand Association constant Elution conditions (step) Sample residence time
HIC	Hydrophobicity	Good resolution Good capacity Good speed	Capture (++) Intermediate (+++) Polishing (+)	High ionic strength No volume limitation	Low ionic strength Concentrated sample	Hydrophobic ligand type Choice of salt concentration Gradient slope
RPC	Lipophilicity	High resolution Low capacity Low speed	Intermediate (+) Polishing (+++)	Limited sample volume (<5% c.v.) and flow rate range	In organic solvent, risk of loss in biological activity	Media backbone Gradient slope of modifier Sample load

A. Basic Parameters

1. Parameters of the Chromatogram

A typical chromatogram of three Gaussian peaks obtained for the separation of three components of a sample is shown in Fig. 3.

The column dead time, t_o, corresponds to the column residence time of a nonretained component, and coincides with the arrival of the solvent front at the end of the column. The column dead time is related to the volumetric flow rate of the mobile phase, F, and the total volume of mobile phase in the column, also called the column dead volume, V_m.

$$t_o = V_m/F. \quad (1)$$

Equation (1) assumes that the time required for the sample to move from the injector to the column inlet and from the column outlet to the detector is negligible. The t_o is used to determine the corrected retention times t'_r of the different sample components from the respective times taken by each retained component to be detected t_r (see Fig. 3).

$$t'_r = t_r - t_o. \quad (2)$$

It is common to use retention volume rather than time in order to be able to compare different conditions of flow rate, or different sized columns. From Eq. (1), we have

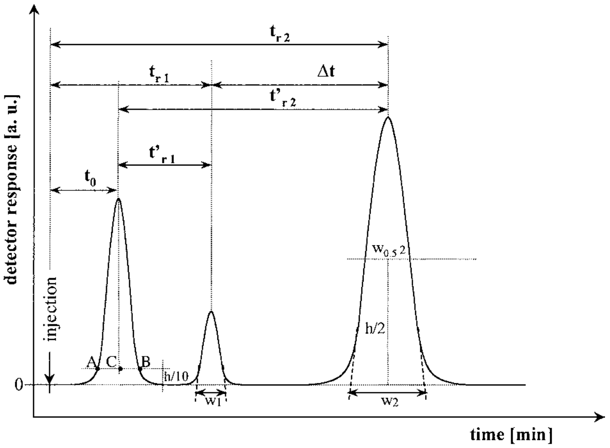
$$V_m = F \cdot t_o, \quad (3)$$

and similarly, the retention time t_r corresponds to a retention volume V_r.

$$V_r = F \cdot t_r. \quad (4)$$

2. Retention

Retention is the basis of chromatographic separation, as it refers to the fact that the different compounds are retained by the column to a varied degree. The phenomenon is quantified by the definition of the retention factor k', which



Retention	$k' = (t_r - t_0)/t_0$
Efficiency	$N = 5.54(t_r/w_{0.5})^2$
Asymmetry	$As = CB/CA$
Selectivity	$\alpha = k'_2/k'_1$
Resolution	$Rs = \Delta t/(0.85(w_{0.5\ 1} + w_{0.5\ 2}))$

FIGURE 3 Example of chromatogram obtained after the separation of three compounds shows three Gaussian peaks. Evaluation of the separation is based on the main chromatographic parameters calculated from graphical data. (*h*: Peak height.)

is given by the amount of solute (in moles) in the stationary phase W_s relative to the one in the mobile phase W_m.

$$k' = W_s / W_m. \quad (5)$$

The retention factor is related to the retention volume of the solute V_r by

$$k' = (V_r - V_m)/V_m. \quad (6)$$

The retention factor can also be expressed in regard to time instead of volume (see Fig. 3). The concept of retention factors was developed for isocratic (from Greek iso = the same and cratos = strength) elution, i.e., under conditions where the composition of the mobile phase does not change throughout the separation. In the case of gradient elution, simple retention times or volumes are used instead.

The k' is related to the distribution coefficient, K_D, expressing the concentration of solute in the stationary phase, C_s, over that in the mobile phase, C_m by

$$k' = K_D \cdot V_s / V_m, \quad (7)$$

where V_s is the volume of the stationary phase. The retention factor can only be constant if the distribution coefficient K_D does not vary. In preparative, i.e., usually nonlinear chromatography, a change in the initial concentration of sample can result in a shift in the retention factor depending on the form of the connected adsorption isotherm function. The retention factor is proportional to the phase ratio ε (i.e., V_s/V_m). V_s may vary with the specific surface of a stationary phase material, even if the apparent column or particle volume is the same.

3. Column Efficiency and Zone Width

Column efficiency refers to the ability of a column to achieve separation of very narrow bands in the final chromatogram (small peak widths). The peak width (at the base), w (=4σ for a Gaussian peak) and correspondingly the variance of the zone, σ^2, are primarily affected by zone broadening effects in the column. In addition, they are proportional to the distance traveled by the zone, z. The zone broadening per unit length is called the plate height and is denoted H (or HETP, height equivalent to a theoretical plate).

$$H = \sigma^2 / z. \quad (8)$$

Setting $z = L$, the length of the column, gives the relationship

$$H = \sigma_L^2 / \mathrm{L}. \quad (9)$$

More commonly, N, the column efficiency or number of plates per column, is determined experimentally from the chromatogram by

$$N = 16 \cdot (t_r / w)^2. \quad (10)$$

Determination of w at the baseline is not convenient and often the width at half-height of the peak is used, $w_{0.5}$ (see Fig. 3).

$$N = 5.54 \cdot (t_r / w_{0.5})^2. \quad (11)$$

Another relationship for N, which is used with many modern data processing systems, is

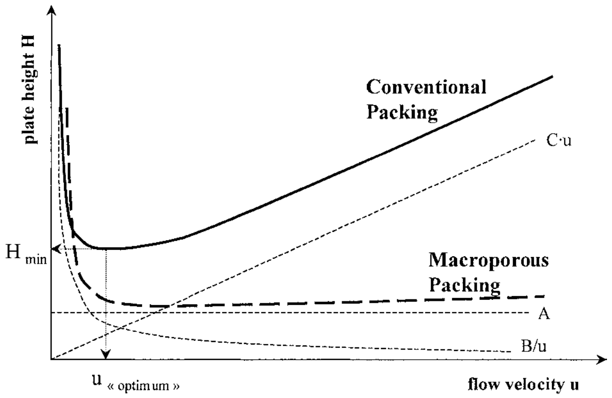
$$N = t_r^2 \cdot h^2 \cdot 2\pi / A^2, \quad (12)$$

in which h represents the peak height and A its area. The aim of optimizing a chromatographic separation is to have a column with the highest possible efficiency, meaning the highest possible number of plates per meter.

The flow velocity, u, of the mobile phase has an important effect on H. The flow velocity is expressed in distance (=column volume/cross-sectional area) per unit of time in contrast to the (volumetric) flow rate F. The so-called van Deemter plot is typically used to describe the change in the plate height H as a function of the mobile phase velocity u (Fig. 4).

$$H = A + B/u + C \cdot u. \quad (13)$$

A is related to eddy diffusion, B to longitudinal molecular diffusion (in the mobile phase), and C to mass-transfer resistance (lateral diffusion in the stationary phase).



Van Deemter equation:	
$H = A + B/u + C \cdot u$	
A	eddy diffusion
B	longitudinal diffusion
C	mass-transfer resistance

FIGURE 4 The efficiency of a column is given by the number of plates (or the plate height). The van Deemter equation relates the plate height H to the mobile phase velocity u. Conventional column packing shows an optimal flow velocity and a decrease in efficiency at both higher and lower flow rates, whereas novel macroporous stationary phases do not lose efficiency with increasing flow rates, thanks to a resistance to mass transfer (C term) close to zero.

For conventional porous stationary phases the van Deemter plot runs through a minimum. At very low flow rates (B-term region), molecular diffusion is the predominant reason for zone spreading. At comparatively high flow rates (C-term region), intraparticular mass transfer resistance and non-equilibria become decisive. The minimum value of H should be around 2–3 times the particle diameters for a well-packed column and the corresponding flow velocity is termed optimum velocity. The study of the van Deemter curve is a good method for determining the optimal flow velocity to be used to achieve highest efficiency and allows optimization of a given separation in terms of various goals (resolution, run time, etc.). The resolving power of a column will be highest when operated at the optimum reduced flow velocity. Due to the low effective diffusion coefficient of high molecular mass molecules, however, this will often be impractical for biopolymers.

4. Peak Asymmetry

Under nonoptimal conditions, separations often yield nonsymmetrical peaks, i.e., other than Gaussian in shape. Asymmetry can be measured in terms of the band asymmetry factor As, that is calculated from measures taken at 10% of the peak height (see Fig. 3) by dividing the width segment after the peak maximum projection and the width segment before the peak maximum projection.

$$As = CB/CA. \tag{14}$$

A "perfect" Gaussian peak is symmetrical and As is hence equal to 1.0. Adequate symmetry is considered achieved as long as bands have asymmetry factors between 0.9 and 1.2. Tailing ($As > 1.5$) and fronting ($As < 0.7$) can cause poor separations, and indicate possible problems of the column or the chromatographic system.

5. Selectivity

The separation quality of two components is strongly affected by the proximity of the adjacent peaks in the chromatogram. An important measure of band proximity is the selectivity, or separation factor, α, given by the ration of retention factors of two adjacent bands (see Fig. 3).

$$\alpha = k'_2/k'_1. \tag{15}$$

For good separations, values of α must usually be at least between 1.05 and 1.10. These values depend on the nature of the solutes as well as on the chemical composition of both stationary and mobile phase. The temperature is also of influence. The concept of selectivity is more relevant in isocratic than in gradient elution.

6. Resolution

Resolution is a measure of the quality of the actual separation achieved between two peaks. In quantitative terms, the resolution between two peaks, Rs, is equal to the difference in retention time, divided by the average of peak widths at the base of the peaks (see Fig. 3). As it is in practice easier to measure widths at half peak height, Rs is usually calculated according to

$$Rs = \Delta t/(0.85 \cdot (w_{0.5^1} + w_{0.5^2})). \tag{16}$$

A resolution of 1.0 corresponds in this context to a nearly baseline separation. If resolution is poor, it is possible to improve it either by (a) improving the efficiency or (b) the selectivity of the column.

B. Separation by Stationary Phase Interaction

Separations in liquid chromatography are based upon differences in the strength of the interaction between the different types of molecules in the sample and the stationary phase. Thus, molecules will be eluted in increasing order of affinity to the chromatographic medium. In order to separate a target biomolecule from the impurities, conditions that favor the interaction of the wanted product with the stationary phase, while decreasing that of the impurities, or vice versa, have to be established. Resolution is achieved by selectively retarding the target component or the impurities to different extents while keeping the dispersion of solute bands as small as possible.

The several types of chromatography exploit differences in simple physicochemical parameters to achieve separation. Examples include the hydrophobicity of the molecules as in hydrophobic interaction chromatography (HIC) and in reversed-phase chromatography (RPC) or their charge density as in ion exchange chromatography (IEC) or to some extent in hydroxyapatite chromatography (HAC). Although the components of a given (protein) mixture will vary in a number of these parameters, chromatographic conditions, e.g., in regard to the choice of stationary and mobile phase, are usually chosen in such a way, that the separation is clearly governed by one of the possible interaction mechanisms.

1. Ion Exchange Chromatography

IEC is at present the most widely applied technique in preparative protein chromatography both at the laboratory and the production scale; the ion exchange principle is shown in Fig. 5.

Compared to most other stationary phases for biochromatography, IEC matrices are characterized by a relatively large binding capacity and hence allow purification

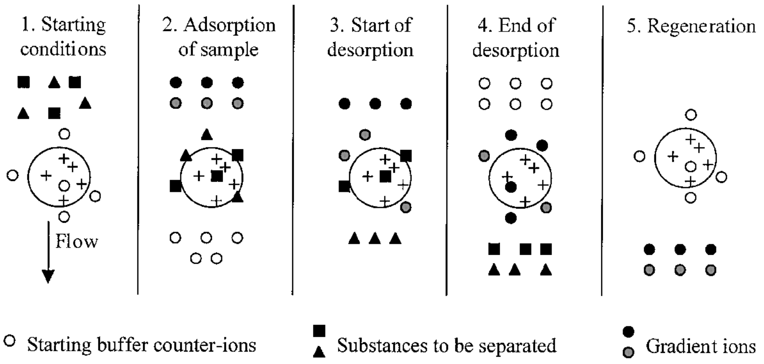


FIGURE 5 Ion exchange chromatography. Molecules interact through their net charge, and are eluted by increasing ionic strength. [From Ion Exchange Chromatography, Principles and Methods. Reproduced with kind permission of Amersham Pharmacia Biotech Limited.]

of comparatively large sample volumes, in comparison to other modes. The recovery of biological activity is usually excellent. Both weak and strong anion and cation exchangers are used. In the case of the weak anion and cations exchangers, usually carrying diethylaminoethyl (DEAE) and carboxymethyl (CM) ligands, respectively, the charge density of the stationary phase surface depends on the pH of the mobile phase. Strong anion and cation exchangers carrying sulfonic acid (S) or quaternary ammonium (Q) groups are independent of the pH in that sense. Since the net charge of the protein, and in the case of the weak ion exchangers, the charge of the chromatographic surface are both pH dependent, control of the mobile phase pH is very important in IEC and great attention has to be paid to the nature of the buffer as well. The sample should be at the same pH as the initial mobile phase, and of comparable ionic strength, in order to maximize the binding.

In IEC, the sample components are retained by virtue of electrostatic interactions between the charged molecules and the oppositely charged chromatographic surface. During binding, the target molecule is concentrated, and subsequently may be eluted in a purified and concentrated form. Elution with (linear or step) gradients of increasing salt concentration (mostly NaCl) is most widely used in the IEC of proteins. The increase in the salt concentration of the eluent results in a "screening" of the charges present at the protein and at the stationary phase surface. As a result the attraction is diminished and the proteins elute. A pH gradient may also be used for protein elution; however, due to the technical difficulties in generating smooth and reproducible pH gradients, this principle is less commonly employed.

By a rule of the thumb, retention occurs in IEC when the sign of the fixed charges at the surface is the opposite of that of the net charge of the protein, which in turn is proportional to the difference between its isoelectric point and pH, provided that the ionic strength of the surrounding buffer is low. However, the relationship between retention and net charge is usually not so straightforward, since the charge distribution over the surface of the protein molecule is not uniform and steric effects also play an important role in determining the magnitude of interaction. Two models, the stoichiometric displacement and the electrostatic interaction model are currently used in protein IEC that link the respective retention factors to the ionic strength of the mobile phase and the number of charged groups involved in the adsorption/desorption process.

2. Hydrophobic Interaction Chromatography

HIC was developed in the 1970s especially for the preparative separation of proteins using predominately hydrophilic, agarose-based stationary phases into which some mildly hydrophobic ligands had been imbedded at fairly low density. Most of these early stationary phases contained in addition ionic groups, retention was therefore due to a mixed mode mechanism. More recently, rigid macroporous silica or polymeric supports have been introduced that are covered with a covalently bound hydrophilic surface layer that incorporates appropriate hydrophobic ligands, such as short alkyl, aryl, or polyether chains at a comparatively low concentration. Protein retention and selectivity depend on the nature and size of the hydrophobic moieties. In practice, the column temperature, the eluent pH, and the nature of the stationary phase matrix will also have a significant influence on a HIC separation. The driving force for retention in HIC is a hydrophobic effect, i.e., less an attraction between the protein molecules and the stationary phase but rather the tendency of the surrounding water molecules to avoid contact with a hydrophobic surface and hence to bring such surfaces into direct contact with each other, as illustrated in Fig. 6.

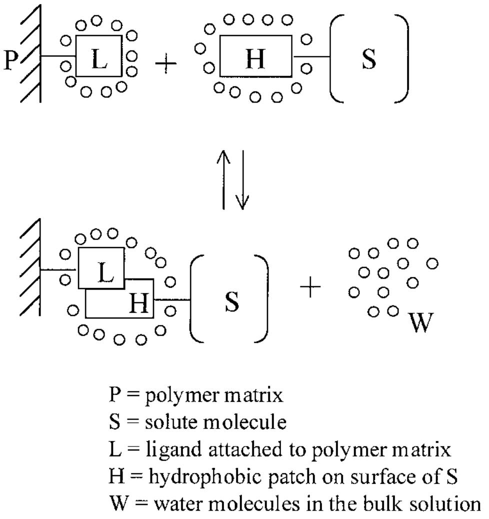


FIGURE 6 Hydrophobic interaction chromatography. The molecules are bound in presence of high salt as a result of increased hydrophobic interaction and eluted by decreasing the salt concentration. [From *Hydrophobic Interation Chromatography, Principles and Methods.* Reproduced by kind permission of Amersham Pharmacia Biotech Limited.]

According to the solvophobic theory of HIC, the magnitude of retention is mainly determined by the so-called cavity term, which expresses the free energy change upon adsorption that is due to the reduction in hydrophobic surface (of the sample molecule or the hydrophobic moiety of the stationary) exposed to water (i.e., the mobile phase) as the two combine. The cavity term is given roughly by the product of the microthermodynamic surface tension of the mobile phase and the molecular contact area upon binding. Thus retention can be considered a solvent effect and generally the retention factor decreases with the surface tension of the mobile phase, i.e., when the salt concentration of a given eluent is lowered. Retention is enhanced by a high salt concentration in the aqueous mobile phase and therefore gradient elution with decreasing salt concentration is most commonly used in protein HIC. Reducing the polarity of the eluent is another way to cause desorption of the proteins—for example, by using an ethyleneglycol gradient. The addition of chaotropic substances (guanidine, urea) or detergents as well as a change in temperature or the pH have been used. The typical initial conditions (e.g., 1.5 M ammonium sulfate or 4 M sodium chloride), make HIC an ideal "next step," e.g., after precipitation with ammonium sulfate or elution in a high salt (NaCl) buffer during IEC.

As in IEC, aqueous mobile phases may be used in HIC. Due to the comparatively mild operating conditions of HIC, the molecular integrity of the native protein is normally well preserved and no significant loss of biological activity occurs. For this reason, HIC is widely employed in preparative and process scale isolation and purification of proteins. It should be noted, however, that some slight changes in the protein structure might occur in HIC due to a weakening of the hydrophobic forces responsible for maintaining that structure.

3. Reversed-Phase Chromatography

RPC has become the predominant branch of analytical chromatography (HPLC) in the life sciences and in biotechnology. Preparative applications of RPC are the exception, although some have been reported. Most commonly, bonded high performance stationary phases prepared by covalently binding hydrophobic ligands such as C_4-, C_8-, and C_{18}-alkyl chains or aromatic functions to the surface of a rigid siliceous or polymeric support are used. Due to the pronounced hydrophobic character of the stationary phase, proteins and peptides bind tightly from a neat aqueous mobile phase and require hydro-organic eluents for release. The separation of peptides and proteins in RPC is therefore typically carried out by gradient elution with increasing concentration of an organic modifier such as acetonitrile, methanol, tetrahydrofuran, and isopropanol. In addition, the mobile phase usually contains low levels of trifluoroacetic or phosphoric acid. The role of the acids is to protonate the residual silanol groups at the surface of the siliceous support and the carboxyl groups of the eluates as well as to form ion pairs with the charged amino groups of the substances to be separated. Ion-pairing agents such as perchlorate can be used at neutral pH.

The retention strength increases roughly with the size and the hydrophobicity of a given substance. In the case of proteins and larger polypeptides, the prediction of the retention behavior becomes difficult, since the three-dimensional (3D) structure, i.e., the number and location of the hydrophobic patches at the molecule surface, becomes decisive. Moreover, under the conditions of RPC with its acidic, hydroorganic mobile phases, many proteins tend to denature and unfold either partially or completely. Oxidation, deamidation, aggregation, and fractionation are also possible. The use of RPC in preparative work commonly requires the refolding of the product into its native configuration after separation. This has been shown to be possible for a number of peptides and smaller proteins, which are of interest to the pharmaceutical industry such as h-insulin and h-growth hormone (hGH). The number of preparative RPC applications in this area remains nevertheless rather small. RPC is, on the other hand, often used for the final polishing of oligonucleotides and peptides made via chemical synthesis.

4. Hydroxyapatite Chromatography

HA is an inorganic material that has been used as the stationary phase for biopolymer chromatography since 1956. The early materials were soft powders (Tiselius apatite), but more recently ceramic HA and also fluoroapatite (FA) beads have also become available, which are much more suitable to the requirements of chromatography in terms of mechanical strength and chemical stability. Both apatites are stable at elevated pH, but will dissolve rapidly below a pH of 5.0.

HA and FA bind both negatively and positively charged substances, yet a simple ion exchange mechanism does not account for the observed chromatographic behavior. Two types of binding sites are present at the chromatographic surface—the calcium ions (C-sites) and the phosphate groups (P-sites)—which are assumed to interact with the amino and carboxylic groups of the protein (see Fig. 7).

In contact with the mobile phase at neutral pH and above, the apatite surface carries a negative net charge as a result of a surplus of phosphate groups. This is amplified in most HA-chromatographic separations on apatite by the use of a phosphate buffer as mobile phase. Positively charged proteins ("basic" proteins) bind by electrostatic interactions to this negatively charged surface. Desorption is brought about by charge screening, i.e., a high salt concentration in the mobile phase. By a different mechanism, but with comparable efficacy, cations with a high affinity to phosphate, such as Ca^{2+} or Mg^{2+}, will displace basic proteins from the HA surface. Negatively charged ("acidic") proteins are taken to bind via their carboxyl groups, which chelate the C-sites at the surface. Consequently, proteins with clusters of carboxylic groups are especially strongly bound. However, acidic proteins are also repelled by the negatively charged apatite surface and therefore retained more weakly than basic proteins under standard conditions. They are readily displaced by phosphate, fluoride, and other anions capable of binding strongly to calcium. The most common protocol in protein separation on HA calls for elution in a gradient of increasing phosphate concentration. In this way, all proteins are eluted: first the acidic proteins due to a specific complexation of the C-sites by the phosphate ions and subsequently the basic proteins due to general charge screening. A rather interesting alternative for downstream processing is the double gradient method, where first the basic proteins are eluted from the surface in a gradient, e.g., of increasing Cl^- concentration, followed by the elution of the acidic ones in a phosphate gradient.

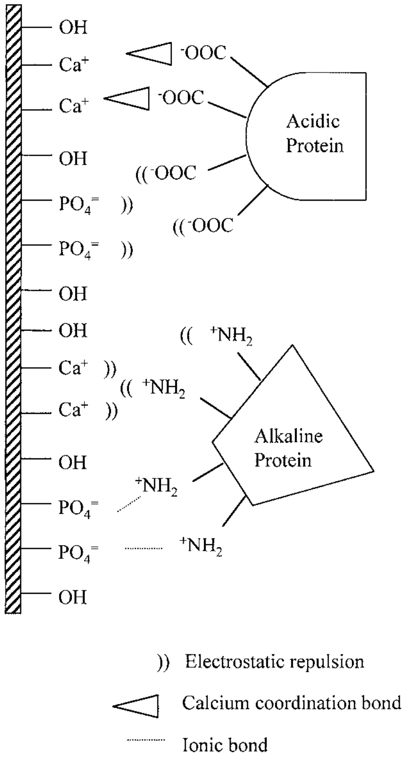


FIGURE 7 Different forces involved in the adsorption of charged proteins on hydroxyapatite.

The different retention mechanisms for acidic, neutral and basic proteins facilitate the development of group separation schemes in HA chromatography according to the isoelectric points of the proteins. In processing products obtained from mammalian cell cultures, HA chromatography has, e.g., been used to separate serum albumin, an acidic protein, from any more basic protein such as Immunoglobulin G (IgG), in a quick and simple fashion. HA chromatography of proteins has been known to offer high selectivity and high resolution. HA chromatography is also used for the purification of nucleic acids.

C. Separation by Affinity (Biospecific Interactions)

Affinity chromatography (AC) exploits biospecific interactions for separation purposes and has become increasingly popular in the last decade because of the unique selectivity of the method. The basis for the interaction in this case is the nearly perfect steric fit between two molecules, the so-called affinity ligand anchored to the stationary phase matrix and the target molecule. If this is possible, a sufficient number of weak, usually noncovalent interactions can be realized, with the overall result of a strong and highly specific retention of the target molecule. However, affinity ligands used in preparative AC should to allow for the elution of the product under conditions, which are not too harsh or even destructive. Under such circumstances, AC is the most adequate means to capture a

high-value target molecule present in very low concentrations, i.e., in the microgram to milligram per liter range, in a complex environment such as a cell culture supernatant. This is, for example, the case for most recombinant therapeutic proteins expressed in mammalian cells. Such cell culture supernatants tend to contain many other proteins in much higher abundance than the target molecule, thus necessitating the use of biospecific interactions to capture the target product. In such cases and on the production scale, AC presents a rather attractive single-step alternative to multistep processes incorporating IEC, HIC/RPC, and SEC separations. Alternatively, AC may be used with the aim of specific removal of a contaminant from the product. The application potential of AC in general is wide. At present, however, the costs of the affinity ligand tend to prevent its use save for the high-value products of molecular biotechnology.

AC is usually performed in the frontal mode under conditions where only the target molecule binds to the stationary phase, while all other feed components move through the bed unretained. The surface of the affinity sorbent should be highly hydrophilic and without functions that promote nonspecific interactions. The most commonly used supports are based on agarose, porous glass, silica, polyacrylamide, methacrylate, and cellulose. Fibrous supports were developed specifically for preparative applications. Nonspecific adsorption can be further reduced by carefully choosing the operating conditions regarding the pH and salt concentration of the mobile phase and the additives used. After the impurities have left the column, the product is eluted in a suitable buffer. Desorption is normally achieved by altering the pH, increasing the salt concentration, or introducing a chaotropic agent. Temperature changes, or reversible denaturation, are also used occasionally to cause desorption.

Typical affinity ligands in AC include antibodies, antigens, lectins, receptors, enzyme inhibitors, hormones, and triazine dyes. Protein A and Protein G are widely used as group-specific ligands for the isolation of antibodies, even though they exhibit some subclass specificity. Antibodies themselves are also powerful ligands in AC, due to the fact that they can be readily raised against most products of biotechnological importance. A lot of research has been invested in the last years to design new and improved affinity ligands, either derived from a natural molecule or constructed *de novo*, e.g., via combinatorial chemistry, with the aim of improving the stability, the specificity, but also the price of the AC stationary phases. Recent advances in genetic engineering have also helped to expand the scope of AC. It is possible to fuse an affinity tag such as a Protein A or a polyhistidine sequence to a recombinant protein, which increases the product affinity for the corresponding affinity column considerably and allows selective removal of the target molecule from most contaminants. The construction of a recombinant protein having a fused tag to facilitate its purification may provide a more economical and safer production process even though the cloning may be more elaborate and the tag need to be removed before the final purification of a pharmaceutical.

The affinity ligand may be covalently linked to the stationary phase surface via a hydroxyl, amino, or carboxyl function. Frequently, a spacer arm is used to anchor the ligand to the matrix in order to improve accessibility. A variety of preactivated stationary phases are commercially available for the convenient attachment of affinity ligands. Immobilization of the affinity ligand tends to lower its affinity for the target molecule compared to the value measured in free in solution often by more than one order of magnitude. Ligand leakage and the concomitant loss in capacity and product contamination are serious problems in affinity chromatography, and so is the fact that many affinity columns will also bind denatured or otherwise malformed product molecules or product fragments.

1. Immobilized Metal Affinity Chromatography

Immobilized metal affinity chromatography (IMAC) has been used to purify albumin, monoclonal antibodies, immunoglobulins, blood factors, interferons, enzymes, and many other proteins and polypeptides. Lately the technique has gained general significance in preparative protein purification since fusion proteins carrying a polyhistidine tag can be created, which have a very high affinity to IMAC columns. Protein IMAC is based on the interaction between a metal ion-based electron acceptor (Lewis acid) anchored to the stationary phase and an electron donor (Lewis base) on the surface of the protein. The proteins in most cases interact through their surface histidine and—to a lesser extent—their tryptophane residues. As mentioned above, recombinant proteins can be expressed by adding an oligohistidine tag and thus impart to them high affinity toward the immobilized metal ions. This approach may have advantages over the use of larger tags such as Protein A, which may be difficult to remove to obtain the correctly folded and biologically active protein product.

Metal ions of intermediate polarizability such as Cu^{2+}, Ni^{2+}, Zn^{2+}, and Co^{2+} are particularly suited for interaction with proteins as they may interact not only with the nitrogen in amino and imino groups, but also with oxygen and sulfur. Metal ions are immobilized on the stationary phase by chelating agents, such as the two-dentate ligand IDA (iminodiacetic acid) bound to the support. The nature of the chelating agent is of consequence. If the immobilization of the metal ion involves several coordination sites, the metal is bound strongly to the chromatographic surface and bleeding is less likely to occur. At the same

time, however, the number of coordination sites available for protein binding by the stationary phase is reduced.

IMAC has been used in the chromatographic isolation of proteins and nucleotides, but also to investigate the surface topography of protein histidine residues. The three-dimensional protein structure is not strongly affected by the binding to the chromatographic surface; therefore the biological activity of the product is usually well preserved. Traditionally, soft agarose beads were used for protein purification by IMAC. Since then a number of rigid supports and chromatographic membranes have become available. The mobile phase in IMAC is a buffered salt solution and the strength of the metal–protein interaction is modulated by the type and concentration of the salt. In preparative IMAC, the separation is often achieved by differential elution with stepwise changes in the salt concentration. The pH also influences the retention behavior of proteins and elution in IMAC may most conveniently be achieved by lowering the mobile phase pH to 6 so that the histidine residues are protonated. An agent competing for the metal sites, e.g., glycine, histidine, or imidazole, or an organic modifier may also be used for the elution of the protein. Complexation and hence removal of the chelated metal ions by EDTA presents another means for protein elution.

2. Chiral Chromatography

Chiral or enantioselective chromatography (chiral-HPLC) may be considered another subdivision of affinity chromatography, which deals with small biochemicals rather than with biopolymers such as proteins. Enantiomers are nonoverlayable mirror images of one another, mostly molecules containing asymmetrical carbon atoms. Two enantiomers have the same physicochemical properties regarding charge, size, and solubility, but may differ considerably in their biochemistry and activity. Even though stereoselective chemical syntheses are often possible, the (large-scale) separation of drug enantiomers in order to produce an enantiopure drug remains a common challenge in the pharmaceutical industry. In some cases the scale of these separations makes even the utilization of a simulated moving bed (see below) feasible. Another area, which requires enantiopure substances, is the study of drug metabolism.

Chiral-HPLC started in the 1980s, and the main applications have been the separation of sugars, amino acids (small peptides), and their derivatives. Different types of natural and modified cyclodextrins (CD) have been immobilized as enantioselective ligands onto conventional silica particles. Cyclodextrins are rings of glucose units, with a toroidal three-dimensional (3D) structure. Their hydrophilic (abundant hydroxyl groups) surface makes them water-soluble, while their tore cavity is rather apolar. The separation mechanism is based on the partitioning of the enantiomers between the outer and the inner surface according to their polarity, and their inclusion in the tore cavity. The chiral retention mechanism hence mixes ionic interaction, hydrogen bonding, dipole interactions, steric, and π-π interactions as well as inclusion complexation. CD phases are used in the polar organic mode or in the reversed-phase mode. Mobile phases for the polar organic mode are mainly acetonitrile and methanol, made polar through the addition of anhydrous acids and bases, whereas the mobile phases for reversed-phase chromatography contain water. The pH, the buffer concentration and type, as well as the organic modifier concentration and type will affect the separation. Immobilized macrocyclic antibiotics (glycopeptides) have also been used as chiral selectors, very much in the same manner as cyclodextrins. Immobilized α_1-acid glycoprotein (AGP) and human serum albumin (HSA) are known to act as chiral selectors, which act via ionic binding, hydrogen bonding and hydrophobic interaction. These ligands are mostly used in the reversed-phase mode.

D. Separation by Size

In size-exclusion chromatography (SEC), also called gel filtration (GF) or gel permeation chromatography (GPC), the sample molecules are separated according to their hydrodynamic diameter (i.e., ultimately according to their size and mass) and not as the result of an interaction-mediated process. The sample is passed though a column packed with an inert porous material having appropriate pore size distribution and volume. In analytical SEC, the sample size should be no more than 3% of the column volume (c.v.); however, in preparative work the sample may occupy up to 15% of the column volume. Since no interaction between the solute and the stationary phase occurs, all the sample components are eluted within one column volume, resulting in a quite low resolving power of complex mixtures by this method. In biochemical applications, SEC is hence mostly used for the rapid and convenient separation of sample components having substantially different molecular masses, such as in the desalting of a protein solution (or buffer exchange). Preparative SEC is often used as a polishing step following other chromatographic separations.

In SEC, separation occurs due to differences in the accessibility of the intraparticular void volume by the sample components of different molecular dimensions. Molecules larger than the upper exclusion limit cannot enter the intraparticular void space and elute first, whereas sufficiently small molecules have access to all the pores, and therefore elute last, see Fig. 8.

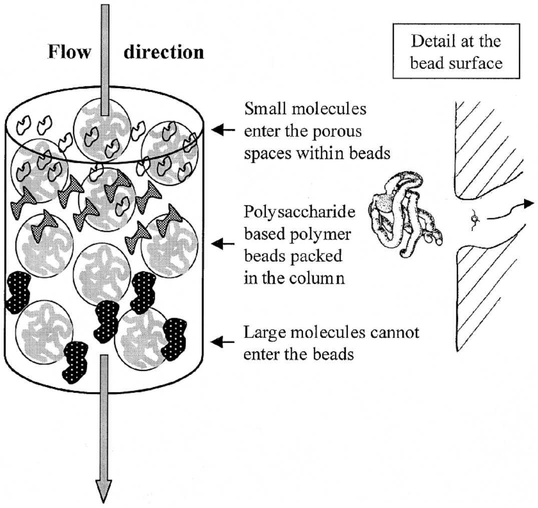


FIGURE 8 Size-exclusion chromatography. Molecules larger than the upper exclusion limit cannot enter the intraparticular void space and elute first, whereas sufficiently small molecules have access to all the pores, and elute later.

For sample components of intermediate molecular dimensions, the retention volume, V_r, is given by

$$V_r = V_m + K_D \cdot V_i, \quad (17)$$

where V_m is the interstitial volume, K_D is the distribution ratio and V_i is the intraparticular void volume.

The magnitude of K_D is determined by the fraction of intraparticular volume, which can be entered by the molecule of interest. Several models have been put forward to relate K_D to the properties of the molecule and the chromatographic matrix. Plots of K_D the logarithm of the molecular mass are generally linear between K_D values of approximately 0.15 and 0.8, and SEC may thus also be used for a rough estimation of the molecular mass of a macromolecule. SEC is commonly carried out with cross-linked macroporous dextran-based beads such as Sephadex from Pharmacia, or modified agarose and polyacrylamide-based gels, i.e., matrices that do not permit the use of high pressures. However, column packings of high mechanical stability, ranging from silica-based materials to macroreticular rigid polymers, are increasingly becoming available for HPLC-SEC.

II. PROCESS DESIGN IN CHROMATOGRAPHY

An efficient purification process for a given biological usually consists of several consecutive steps, comprising a combination of the different chromatographic modes described above. Only a sequence of separations according to different separation principles can assure that the required purity levels of sometimes more than 99% are reached. In some cases, e.g., in the case of assured virus removal, a minimum number of steps with a certain performance (measured in orders of magnitude or "logs") is required by the authorities. In a typical downstream process the techniques are chosen in such an order as to produce a good selectivity but also recovery for the target product, in function of its surface properties, and to remove all impurities below the level of acceptance. The intricacy of the process development is to optimize the downstream process in such a way that both purification and quality prove to be satisfactory. In order to meet the needs of industrial biotechnological processes, purification methods should be fast and highly specific, and highly cost-effective. The primary goal is thus finding a suitable combination of different modes that present a balanced compromise between yield and purity of the target biomolecule in as few steps (usually <5) as possible using complementary separation mechanisms. The following sequences are often found and can be considered "classical": IEC followed by HIC (good interface, since the separation principles of "charge" and "hydrophobicity" are orthogonal and the high salt elution buffer typical in IEC is an excellent loading buffer for HIC) followed by GF for polishing (e.g., removal of aggregates). Another powerful sequence is presented by AC followed by IEC and finally GF.

Among other things, the adopted strategy is dictated by the composition of the feed. Thus, IEC can be suitable for initial cleanup form crude broth, and AC can result in an initial high purification *cum* concentration provided fouling can be avoided. Within their respective limits, both HIC and RPC are powerful techniques for the separation of closely related molecules (e.g., the target molecule and some of its degradation products). GF may be used for the final polishing step with the aim of removing impurities and contaminants differing from the product mainly in size (e.g., product monomers from dimers). Tables II and III detail for each chromatographic mode the suitable areas of application, i.e., initial capture, intermediate purification or final polishing, with considerations on mobile and stationary phases in Table II, and on sample characteristics in Table III. The most important parameters for the optimization of each mode in regard to its successful application in preparative chromatography are summarized in Table III.

Chromatographic separations are traditionally batch procedures. A certain volume of the mixture to be processed ("sample," "feed") is introduced at one end of a column packed with the stationary phase (conventionally: porous particles) through which the mobile phase is passed. The chromatographic separation of the sample components takes place along the axis of the column and results from the differences in physical and/or chemical

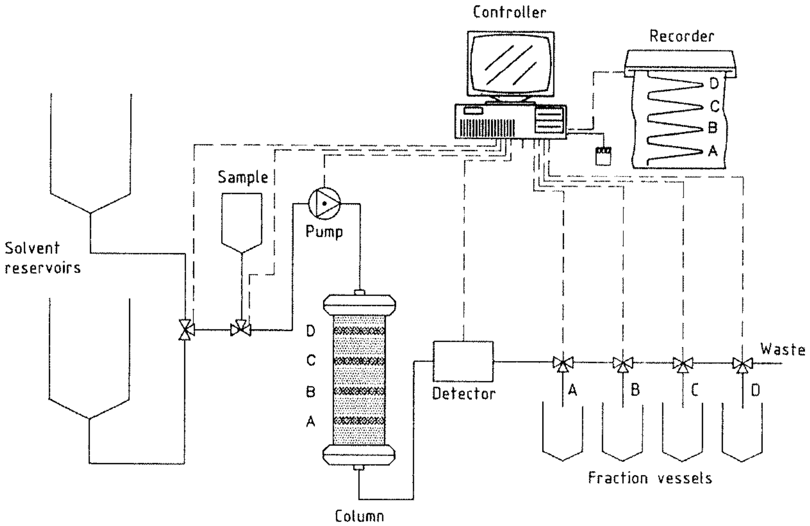


FIGURE 9 A chromatographic setup including buffer vessels and valve, pumps, mixer, a sample feed port, the column, and a detector. The system can be controlled with adequate software; fractions can be collected after detection. [From Freitag, R. and Giovannini, R. (2001). *Biotechnology and Bioengineering*, Wiley-Liss, Inc. a subsidiary of John Wiley & Sons, Inc. With permission.]

interactions between the sample components and the chromatographic matrix. A standard setup for liquid chromatography consisting of buffer reservoir(s) and pumps, a mixer, a sample feeding loop (or pump), the column, and a detector is shown in Fig. 9. A fraction collector is also useful, especially in preparative chromatography.

The pressure drop over the column may be calculated from the Hagen–Poiseuille equation, in which the void fraction (or porosity) of the column, the velocity and the viscosity of the mobile phase, and the particle size are considered. Changing the particle size from 100 to 10 μm improves the plate number, but also increases the flow resistance by two orders of magnitude. The consequent need for high-pressure systems for small particles can be a disadvantage in a production environment, where low-pressure systems are often preferred. The viscosity is of special significance when applying viscous samples (e.g., certain culture supernatants, or solutions of high DNA contents) and when transferring methods to the cold room.

A. The Different Modes of Chromatography

Adsorption (interactive) chromatography as opposed to size-exclusion chromatography is based on the differential distribution of each substance between the stationary and the mobile phase due to interactions between the components and the chromatographic surface. The equilibrium relationship for the distribution of the components between the mobile and stationary phases is given by the respective adsorption isotherms. In the case of biochromatography, the Langmuir isotherm model is widely used, even though particularly biomacromolecules show some deviations from it, e.g., because of multipoint interaction, aggregation or multiple retention mechanisms, and other secondary equilibria. At low sample concentration, Henry's law usually holds and the relationship between C_m and C_s is linear. At higher concentrations, the isotherm becomes nonlinear. Consequently, competition between the various molecules for the adsorption sites takes place and the multicomponent isotherm of a given compound will be suppressed as compared to the single-component isotherm. Taking the isotherm into account, a chromatographic separation can be mathematically described by solving the mass balance equation within the appropriate boundary conditions. According to the inlet boundary conditions, three modes of chromatography can be distinguished—namely, elution, frontal, and displacement chromatography.

1. Gradient and Isocratic Elution Chromatography

Linear elution chromatography (see Fig. 10a) is the preferred operational mode in analytical chromatography. The sample is introduced into the column approximately as a Dirac pulse. The components move through the column with different velocities, due to differences in their distribution between the mobile and the stationary phase, and thus are separated. If the composition of the mobile phase does not change throughout the separation, the process is referred to as isocratic elution. However, many biologicals tend to show an all-or-nothing type of binding under these circumstances, i.e., except for a narrow window of eluent strength, the molecules will either bind

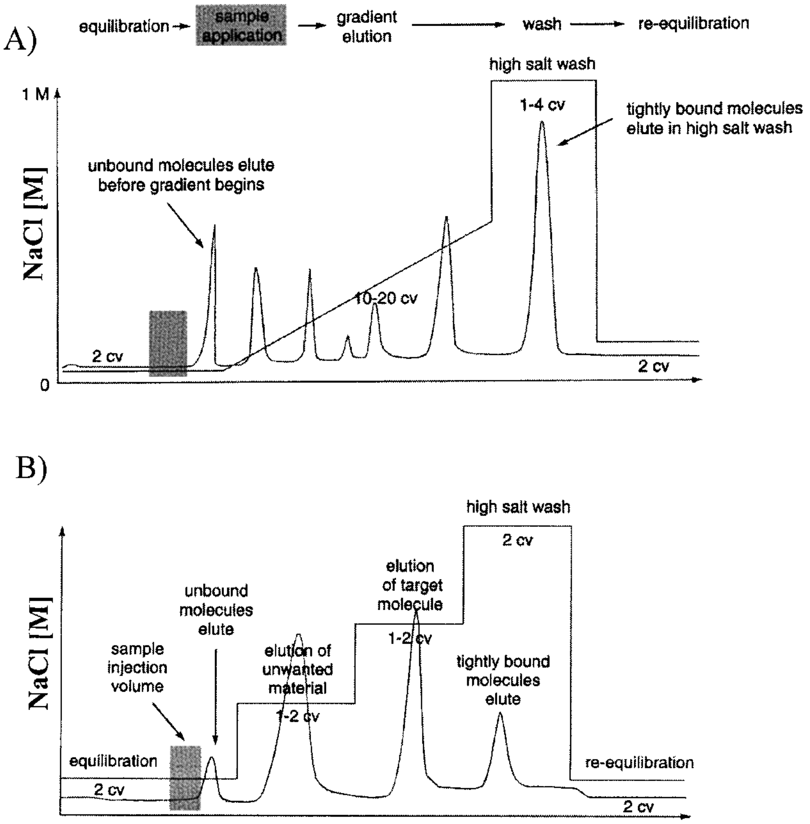


FIGURE 10a Separation of a mixture of biomolecules achieved in a linear (A) or stepwise (B) gradient. [From *Protein Purification Handbook*. Reproduced by kind permission of Amersham Pharmacia Biotech Limited.]

strongly to the surface of the stationary phase or show no tendency to bind at all. Since the elution windows of all sample components will rarely overlap, their separation by isocratic elution is often not practical. Instead, the eluent strength is increased gradually or stepwise (differential elution) during the chromatographic run, to bring about the separation of a multicomponent mixture. If the eluent strength is increased gradually, the migration velocity of the eluates increases with the eluent strength, until they move with the velocity of the mobile phase. Since the molecules at the rear of a peak are moving in a zone of higher eluent strength than those at the peak maximum, they are sped up in relation to the bulk sample, while the opposite effect is operative at the front of each sample zone. Therefore elution with an appropriate gradient involves focusing, and results in relatively sharp peaks and a reduction of peak "tailing." At the same time, gradient elution is much more demanding than isocratic elution as far as instrumentation and theoretical treatment of the process are concerned. Step gradient elution is more commonly used than linear gradient elution in preparative/process chromatography, but also in affinity chromatography. Ideally, two well-chosen elution steps suffice, one for recovering all components that bind less strongly to the chromatographic surface than the product, and the other eluting exclusively the product.

2. Frontal Chromatography

Frontal chromatography is a binary separation process in which only the least-retained component is isolated from the others. The mixture to be separated is fed continuously into the column under conditions that favor the binding of all the components but one. This component is obtained in pure form at the column outlet, until the dynamic capacity of the stationary phase is exhausted and the other sample components break through; a typical chromatogram is shown in Fig. 10b. Frontal chromatography is the first step in many biopolymer purification schemes involving differential elution. The method per se is applicable when the product to be purified has much lower affinity for the stationary phase than the other feed components and therefore breaks through far ahead of the impurities.

3. Displacement Chromatography

This nonlinear multicomponent separation technique is eminently suitable for preparative/process scale applications. In displacement chromatography, the competition

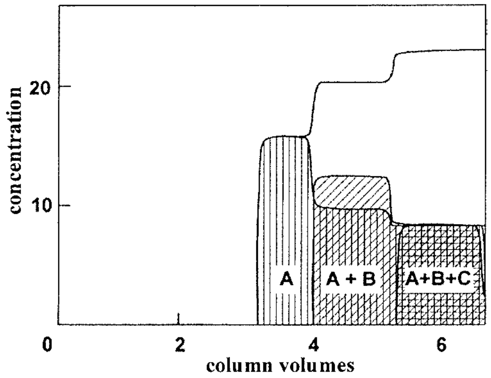


FIGURE 10b Separation of a ternary (1:1:1) mixture by frontal chromatography. The corresponding single component isotherms are assumed to follow the Langmuir equation. [From Antia, F. D. and Horvath, C. S. (1989). *Ber. Busenges. Phys. Chem.* **93,** 963. With permission.]

of the feed components for the binding sites is exploited to bring about the separation. Although the principles of displacement chromatography have been known for more than 50 years, the development of highly efficient HPLC instruments and columns, together with an improved understanding of the theory of nonlinear chromatography, have recently provided new impetus for the use of this chromatographic mode.

In displacement chromatography (see Fig. 10c), the feed is loaded onto the column under conditions that allow for strong binding of all sample components to the stationary phase. Afterward the displacer, a substance with extremely high affinity for the stationary phase, is introduced. As the displacer front advances along the column, the components are forced to compete for the adsorption sites and—at least for systems showing approximately Langmuirian type isotherms—are finally separated into adjacent rectangular bands, if the column is sufficiently long. At this point all bands move with the speed of the displacer front. The concentration of each zone is determined by the isotherms

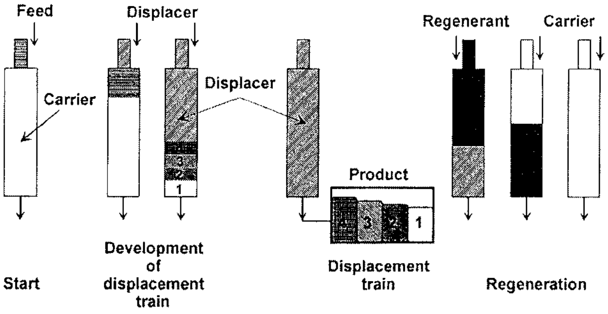


FIGURE 10c Stages of displacement chromatography. [From Horvath, C. S. *et al.* (1981). *J. Chromatography*, **218,** 367. With permission.]

of the respective substances in relation to the isotherm and concentration of the displacer. The concentration of the components may thus be increased in respect to their concentration in the feed. This feature is of interest not only in preparative scale separations, but also for enrichment of certain components in trace analyses. The effect of feed and displacer concentrations, bed length, and mass-transfer effects on the separation have been treated theoretically, and in many cases the results have experimental support.

Although further work is needed to exploit the full potential of displacement biochromatography, results accumulated over the past few years have demonstrated that the technique can be a powerful tool for the purification of antibiotics, peptides, and even proteins. Progress in ion exchange displacement chromatography (IEDC) has been especially fast in recent years. IEDC has been used to separate cephalosporin C from culture supernatant, to isolate alkaline phosphatase enriched in the periplasm of *Escherichia coli*, as well as to isolate the IgG fraction from blood plasma and monoclonal antibodies from ascites. Guinea pig serum proteins and mouse liver cytosol proteins have been isolated by anion exchange displacement chromatography. Recently, recombinant human antithrombin III has been purified from culture supernatants of Chinese hamster ovary cells. In analytical biotechnology, tryptic digests were characterized by the tandem use of high performance displacement chromatography and mass spectrometry. In 1978, Torres and Peterson *et al.* started to develop and later optimized a system using carboxymethyldextrans as displacers. Other IEDC protein displacers include chondroitin sulfate, carboxymethyl starch, and the polycationic polymer polyethylenimide (PEI). Heparin, protamine, block methacrylic polyampholytes, and polyelectrolytes such as poly(diallyldimethylammonium chloride) (PDADMAC) and polyvinylsulfonic acids have also been identified as powerful protein displacers in IEDC. The steric mass action model has been developed by Cramer *et al.* for the simulation of such separations.

B. Scale-Up Considerations

High performance preparative chromatography, if properly designed and optimized, is a competitive industrial purification process. High column efficiencies and fast flow rates permit us to achieve difficult separations in a short time, thus reducing the danger of product degradation during separation and quite often the purification costs as well. If large quantities of a given biological are needed, a most obvious solution is to repeat the chromatographic separation as often as necessary and to pool the obtained fractions. However, this approach is usually

acceptable only for research purposes and will not be considered for large-scale production. In a production environment the separation (column) has to be scaled up. The scaling up of chromatographic separation processes admittedly remains a challenging problem for bioseparation engineering. However, columns with inner diameters of 3.6 m and heights of 12 m have been built and operated, mainly for separations in the oil and sugar industries and more recently also for the large-scale separation of biologicals such as recombinant human insulin and blood plasma fractionation. Sephadex gel filtration columns with inner diameters of up to 180 cm (custom built by Pharmacia, Sweden) have been used to separate milk proteins, amino acids, technical enzymes, and penicillins. Three parameters are usually considered for the optimization of preparative chromatography—namely, yield, purity, and throughput. In preparative and process scale chromatography the throughput, or amount of sufficiently pure product obtained per unit time and column volume, is of major importance.

Linear elution chromatography does not use the full capacity of either the stationary or the mobile phase. Whenever the sample concentrations are in the nonlinear range of the respective adsorption isotherms, the sample molecules compete for the binding sites on the chromatographic surface and interfere with each other's migration. In the case of a single component Langmuirian isotherm, the zone will have an increasingly sharp front and a diffuse rear boundary or "tail" due to the self-interference of the molecules (triangular zones). A simple procedure to maximize the throughput in elution chromatography ("overloaded" elution chromatography), while still maintaining baseline separation, is the so-called touching-band optimization. The sample load is increased until the second component just touches the rear of the first. If the column is severely overloaded, separation becomes largely a matter of the sample composition. In certain cases, the competition of the two components will cause the zone of the less strongly bound component to be pushed ahead by the zone of the more strongly bound one; "sample displacement" occurs often with some felicitous results on the separation. The opposite, a "tag-along" effect, may also be observed. Especially in the case of smaller and chemically/mechanically more stable molecules, a solution to the optimization of the throughput may be to operate the column under conditions of pronounced overloading and simply recycle the mixed zone, i.e., add it to the fresh feed.

To scale up the column, the diameter may be increased, and higher flow rates used. The limitations to this approach are usually technical; for instance, even though the separation factor is large enough to allow for a decrease in separation time, the pressure drop over the packed bed at elevated mobile phase velocities may exceed the pressure rating of the pump or of the chromatographic medium. In that case a decrease in column length may be a better solution. The injection of the sample and the detector capacity may also have their limits. In some cases the problem of combining high flow rates with acceptable pressure drops can be circumvented by using novel stationary phases characterized by lower mass transfer resistance and low backpressure (see also Section III, Stationary Phases for Biochromatography). When scaling up the chromatographic process, the quality of the separation ought to be maintained, which means in the simplest case that the column efficiency, i.e., the number of plates, should be about the same. This can be accomplished relatively easily when only the diameter of the column is increased and the column length and the stationary phase material remain the same. Whereas the mobile phase composition and linear flow velocity are kept constant, the volumetric flow rate of the eluent and the sample load may be increased in proportion to the cross-sectional area of the column. Columns having inner diameters greater than 5 cm should be equipped with some sort of compression device to maintain bed stability throughout the separation. Several designs, for the static or dynamic compression of the column packing, have been patented; dynamic axial compression by a piston being currently perhaps the most suitable method. Dynamic compression is generally preferred to static compression, as the former allows for the adjustment of the applied pressure as the particles swell or shrink. Resolution and peak shape should be independent of the column, keeping everything else constant. In practice, however, a certain decrease in efficiency should be expected because of difficulties in attaining uniform column packing at large scale, due to transcolumn nonuniformities and the consecutive irregular flow distribution. Radial temperature gradients may also be produced with concomitant decrease in column efficiency.

Evidently, a dynamic similarity must be maintained between the functions of the smaller column for which the separation has been developed and optimized, and the larger column to be used for production. Most experience has been gained in scaling up elution chromatography on the basis of similitude relationships. Some data for the design of the large-scale column, for example in regard to the flow rate, the sample load, and the gradient conditions, are available in the literature. For the calculations of production rates and column utilization, the expressions thus derived can be consulted. The length-to-diameter ratio of the large-scale columns will often be smaller and the stationary phase particles larger in preparative than in analytical liquid chromatography. The choice of the particle diameter of the stationary phase will be influenced by most operating parameters when the length of the bed is changed upon scaling up, or the specifications for the throughput

and the desired final product quality are altered. The effect of the particle diameter on the efficiency of the column tends to be less pronounced under overloading conditions. However, an optimal ratio of column length to the square of the particle diameter can be found so that in the most common applications, the optimum particle diameter is in the range from 10 to 30 μm. Broad particle size distribution often leads to an inferior separation efficiency with concomitantly higher backpressure.

C. Processes for Preparative Chromatography

The use of a typical—but large—batch column packed with porous particles is possible and common practice in preparative chromatography; and columns of more than a cubic meter have been used. Some of the problems encountered with biochromatography on preparative columns have been discussed in Section IIB (Scale-Up Considerations). However, several alternatives have been suggested, which either circumvent some of the typical problems encountered when using columns early on in the downstream process or during scale-up (fluidized bed, radial chromatography), or which allow to operate a chromatographic separation in a continuous manner (continuous annular chromatography, simulated moving bed chromatography). On the production scale, a continuous separation process is often preferable, as such processes are usually more economical, are automated more easily, and provide a more homogeneous product quality. Furthermore, they enable a better use of the adsorbent and mobile phase and may facilitate recycling.

The advantages of continuous chromatographic separations are fully realized only in large-scale processes; therefore such chromatographic processes are found mainly in the petrochemical and sugar industries. However, as industrial biotechnology advances, continuous chromatographic systems are expected to play an increasing role also for large-scale purification and isolation of biochemicals, e.g., in the food or (bio)pharmaceutical industry. Such systems can work at countercurrent or crosscurrent flow. The moving bed and the simulated moving bed are examples of the countercurrent approach, since the adsorbent and the eluent move in opposite directions. Continuous annular chromatography, on the other hand, is an example for a crosscurrent flow system. While countercurrent methods are restricted to the separation of binary mixtures, crosscurrent methods are able to also separate multicomponent mixtures.

1. Expanded (Fluidized) Bed Chromatography

In an expanded or fluidized bed, the adsorbent particles are placed in a vessel with a porous bottom plate. A fluid flows

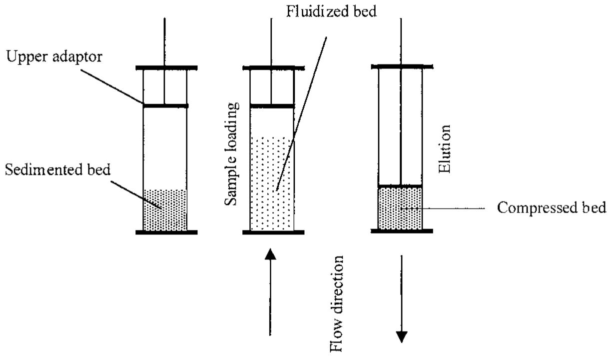


FIGURE 11 Schematic representation of a fluidized bed system.

upward through the porous plate at such a flow rate that the particles become "fluidized" within the confinements of the container, as illustrated in Fig. 11. There are several ways of using adsorptive separations with fluidized beds in biotechnology. A major advantage of the approach is that particles (e.g., also residual cells) pose no problem. As long as they do not stick to the particles, solids move with the fluidizing liquid through the expanded bed and are thus removed without any risk of column clogging, while the product is bound to the adsorbent. The product is simultaneously concentrated and, to a certain extent, purified in a single-step operation. Since solids can be exchanged and recycled in fluidized beds, and clogging is not a serious problem, such beds are putatively interesting systems for applications involving applying the crude cell culture suspension directly into the expanded bed system, thus bypassing the tedious step of cell and cell debris removal.

Since the 1970s, the stabilization of a fluidized bed of magnetic particles by means of applying a magnetic field has been investigated. This was shown to suppress particle circulation to a large extent and solids seem to move in nearly plug-flow manner in such fluidized beds. As both the solid and the fluid movement are controllable, magnetically stabilized fluidized beds are interesting albeit mechanically complex continuous chromatographic separators. Both crossflow fluidized beds, in which the solids move perpendicularly to the direction of the fluid, and countercurrent fluidized beds have been described. Another way to stabilize the fluidized bed is to use adsorbent particles with different densities (establishment of a density gradient). In this case, the heavier particles will stay close to the bottom of the column, while the lighter ones raise preferably to the top. The result is a putative multistage separation system (expanded bed) as opposed to an indiscriminately fluidized bed, which can only achieve a single stage separation.

One can mention that affinity beads are also being used in batch affinity purification by suspending the beads in

the medium, which can contain particles or cells, in order to bind the target molecule in a one-step adsorption process. The beads have then to be separated from the medium by gravity or by applying a magnetic field in the case of magnetic bead cores, or by filtration. Desorption and collection of the target molecule is done with the same eluents as in chromatography. Equal recovery might require slightly higher affinity constants for a single contact in suspension vs multiple contacts occurring in biochromatography.

2. Radial Chromatography

In radial flow chromatography, the adsorbent is packed between two concentric cylindrical porous frits. Eluent and feed flow via capillary channels from the top of the instrument to the outer cylinders through the stationary phase to the inner porous cylinder and then to the exit port of the column, see Fig. 12. The separation path length is therefore comparatively short, i.e., equal to the thickness of the annulus. Radial chromatography has been developed to overcome backpressure problems that can arise when a process is scaled up from the laboratory to the production scale. A unit is easily scalable because the length of the cylinder may be increased while the sample path/backpressure stays constant. Column backpressures are typically very low and the system can be operated at very high flow rates, which may be as high as one or two column volumes per minute. Several applications in the biotechnological field have been reported even at very early stages of the purification process. Another advantage of this type of column is that they are quite easy to pack homogeneously.

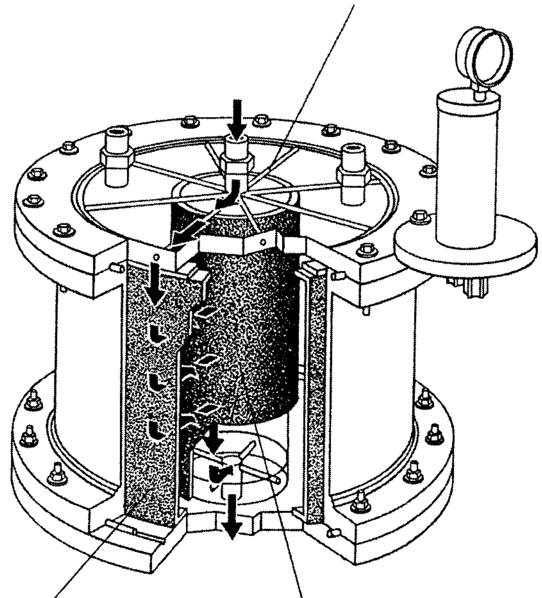


FIGURE 12 Schematic representation of a radial flow column. [From *Bioseparation and Bioprocessing: A Handbook* (1998). Vol. 1, p. 147. Wiley-VCH Verlag, Weinheim. With permission.]

3. Preparative-Continuous Annular Chromatography

The concept of the continuous annular chromatography (CAC) was introduced by Martin in 1949. The packed bed of the adsorbent occupies the annular space between two coaxial cylinders. The bed rotates past a fixed port through which sample is continuously fed. The eluent percolates downward through the bed. Substances elute as helical bands due to the simultaneous axial chromatographic process and bed rotation as shown in Fig. 13. The stronger the adsorption of a given substance is, the further away

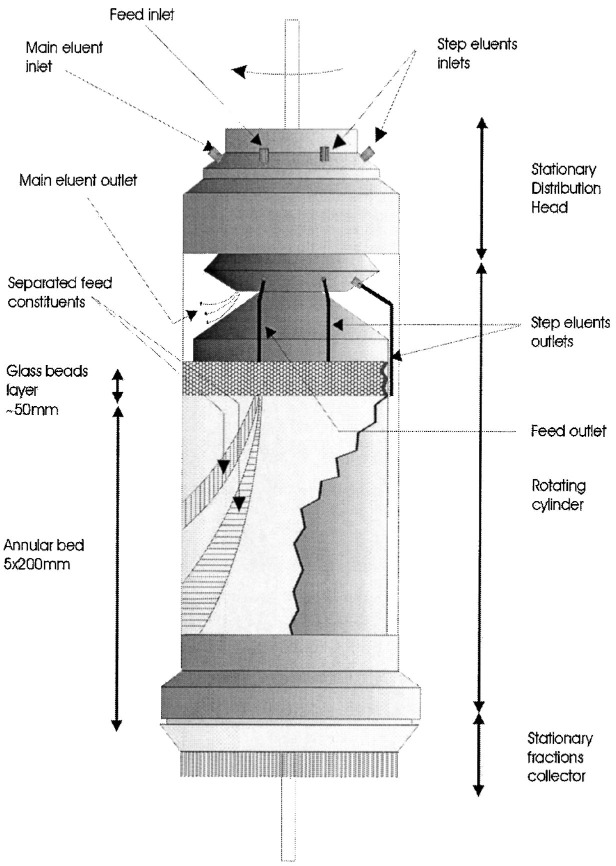


FIGURE 13 Continuous annular chromatograph. [From Freitag, R. and Giovannini, R. (2001). *Biotechnology and Bioengineering*, Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc. With permission.]

from the point of sample introduction it will appear at the bottom of the bed. A relation between a fixed bed process and the same process applied to the CAC is given by

$$\alpha = \omega \cdot t, \tag{18}$$

where α is the elution angle from the feeding point, ω is the rotation speed, and t is the elution time in fixed bed chromatography. An alternative design has been reported where the sample inlet and the collection ports are rotated while the annular bed is kept fixed in order to overcome mechanical difficulties.

Initially, mostly isocratic separations were carried out, and SEC mode was common. However, modern CAC systems also allow for step gradient elution, which is usually done by injecting the respective eluents at fixed positions around the circumference of the bed. Continuous separations of amino acids, sugars, proteins, and DNA have been reported with such systems.

4. Simulated Moving Bed

In the true moving bed system, feed inlet and collection outlets are fixed while the adsorbent moves in the opposite direction of the eluent (countercurrent flow). The system is extremely difficult to operate because it involves the circulation of a solid adsorbent. However, the moving bed can be simulated by a number of columns connected so that by using the appropriate valving system, the operation of a countercurrent is approached by the appropriate shift of the injection and collection points of the columns as shown in Fig. 14. The adsorbent is still moving with respect of the inlet/outlets points if they are moved step by step between a given number of fixed columns. Whereas such systems are not capable of resolving multicomponent mixtures, they offer a promising continuous high throughput alternative for the processing of binary mixtures. Such systems have been used for the purification of dextrans, carbohydrates, and sugars using size-exclusion and ion exchange chromatography mainly in the petrochemistry or in the fine chemicals industry. More recently, the simulated moving bed (SMB) has also a been proposed as a very elegant solution for the efficient separation of racemic mixtures into enantiopure drugs at large scale.

III. STATIONARY PHASES FOR BIOCHROMATOGRAPHY

Traditionally, chromatographic separations of bio-(macro)molecules were fairly slow processes. The typical columns packed with porous particles could not be operated at high speed, due to considerable diffusional constraints. The mass transfer within the pores of the particles, where most of the adsorptive surface is located, occurs mainly by molecular diffusion. The inherently low diffusion rate of macromolecules gives rise to considerable band broadening and a rapid reduction in resolution when increasing eluent velocities are used (see van Deemter curve in Fig. 4). As a consequence, the target molecules are exposed for some time to potentially harmful conditions and significant degradation may occur.

Besides the mass transfer phenomena, some other more technical factors such as limited mechanical stability of the column and the stationary phase itself have been known to limit the speed of chromatographic separations. Different approaches to overcome this problem include the improvement of the conventional porous particles (nonporous, tentacle, gigaporous and hyperdiffusive particles) as well as the introduction of new nonparticle-based stationary phases (membrane, disk, and continuous bed adsorbers). These modifications are directed at speeding up the separations, which have as consequences, e.g., the reduction of losses and degradation by shorter contact with certain solvents as well as the reduction of solvent consumption and manufacturing costs.

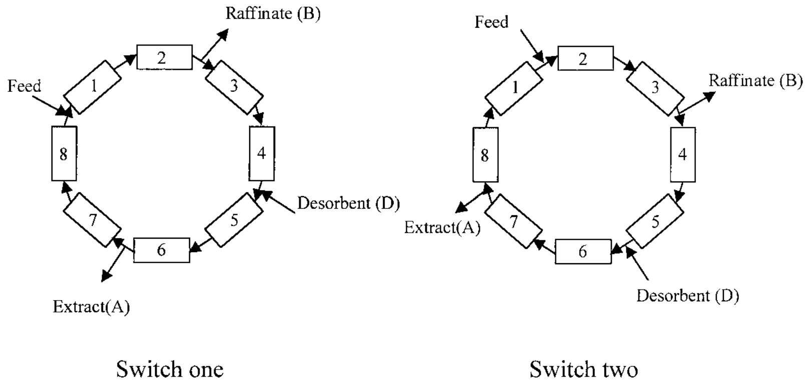


FIGURE 14 Schematic representation of a simulated moving bed (SMB) system with sequential operation.

A. Particle-Based Stationary Phases

1. Conventional Porous Particles

The choice of conventional particle-based stationary phases for (bio)chromatography is vast and details cannot be given here. Table IV attempts to group the main types of stationary phases into categories, such as inorganic and organic (polymeric) materials, which may be of natural or (semi)synthetic nature. The final choice of a stationary phase is governed as much by personal preference and past experience as by considerations such as

- The characteristics of the target molecule (size, isoelectric point, hydrophobicity, possible biospecific interaction)
- The chosen chromatographic mode (stability of the stationary phase toward the mobile phase, required cleaning procedures)

In this context the following advantages and limitation have been noted for the different materials. Conventional silica is sensitive to elevated pH, making sanitizing with alkaline solutions difficult. Polymer-based synthetic supports are more stable in this regard. Silica is, on the other hand, superior in terms of mechanical stability to most if not all other currently existing stationary phases. Polysaccharide-based supports generally show poor mechanical stability; therefore low pressure and low flow rates are indicated, and their scale-up potential is limited. These stationary phases do, on the other hand, show excellent biocompatibility due to their pronounced hydrophilicity. More recently, highly cross-linked polysaccharide gels have been introduced, which are mechanically stable, with a rigidity almost equal to that of silica-based materials. Rigid porous microparticles (average diameter between 2 and 10 μm) with narrow particle size distribution yield very high column efficiencies, but need high pressure at the inlet to reach reasonable flow rates. While such small particles are eminently suitable for analytical separations, bigger particles are usually preferred for preparative separations, because of economic or operating (pressure-drop) considerations. Many preparative applications use columns packed with irregularly shaped particles showing a comparatively broad particle size distribution.

2. Particles with Reduced Mass Transfer Limitation

Several possibilities have been proposed to improve the performance of porous particle-based stationary phases. Most of these approaches attempt to reduce the problem of intraparticular mass transfer and the related loss in column efficiency at higher flow rates. However, due to the high

TABLE IV Particulate Stationary Phases[a]

	Inorganic	Organic
Natural	Chalk	Polysaccharide-based materials
	Charcoal	Cellulose
	Kieselguhr	Starch
	Glass	Agarose
	Sand	Dextran
	Aluminium oxide	
	Magnesium silicates	
	Hydroxyapatite	
Synthetic	Silica beads and monoliths	Methacrylates
	Glass beads of controlled shape, size, and porosity	Polyacrylamides
		Polystyrenes
		Polyamides
		Modified cellulose and various copolymers of controlled shape, size, and porosity

The surface of most of the synthetic supports can be modified, functionalized for each chromatographic mode. For example:
- A hydrophilic layer is added for SEC
- A hydrophobic layer is added for HIC and RPC: C_2, C_4, C_6, C_8, C_{12}, C_{18}
- Ionic groups are added for IEC: strong (QA, SO_3) and weak (DEAE, CM)
- Common affinity ligands such as Protein A and G, lectins

[a] Beads are made of a core material, and their surface is modified accordingly for each chromatographic mode.

price, their application will most likely be limited to the isolation of high-value biologicals.

A very simple way to achieve a reduction in the intraparticular mass transfer is the use of micropellicular, i.e., nonporous particles. However, due to the low capacity of such particles (90% of the adsorptive surface is usually found inside a particle), such nonporous particles are more suitable for analytical than for preparative applications. Another way to increase capacity without relying on the intraparticular surface area, which is more suitable to preparative applications, is represented by the so-called tentacle gels. Gigaporous stationary ("perfusion") particles and hyperdiffusive ("gel in a shell") particles can also be envisaged for preparative separations.

The pores of the hyperdiffusive particles are filled with a gel, which considerably improves the efficiency of the mass transfer by diffusion in such gels. Just as the hyperdiffusive particles, the gigaporous particles were developed in the 1980s to allow the use of higher flow rates and improve the productivity of bioseparation. They contain large through-pores ("convective pores") of several hundred micrometers in diameter, which reach across the particles, together with relatively small and shallow pores ("diffusive pores"), which line the walls of the gigapores and serve mainly to increase the adsorptive surface. When these particles are packed into a column, the mobile phase flows around the particles as in conventional columns but also through the particle. The latter effect reduces the problem of intraparticular mass transfer. However, theoretical considerations have shown that the full advantage of convective mass transfer through the particle is only obtained at relatively high mobile phase flow rates.

B. Continuous Stationary Phases

Continuous stationary phases are single porous entities ("filters," "sponges"), which are transfused by the mobile phase. The walls of the pores through which the mobile phase flows are themselves lined with the interactive surface; the mass transfer resistance of the system is thus reduced to a minimum. Continuous stationary phases have been realized in the form of flat sheets (membrane/disk chromatography), but also in the form of porous rods (continuous bed column chromatography). The van Deemter curve of the various continuous bed-type stationary phases is relatively "flat" even at elevated flow rates, see Fig. 4. In other words, the efficiency of the macroporous material does not decrease by increasing flow rates and such materials can therefore be used for fast bioseparation. The scale-up of continuous beds remains a problem, since the *in situ* polymerization cannot be handled at elevated column scale. Some tube-like columns have been designed for radial chromatography, thus benefiting from the possibility to scale up the separation by modifying the column's length without changing the separation distance.

1. Membranes

Filter membranes are ubiquitous in bioseparation and mainly used to roughly separate molecules according to differences in size. However, affinity, ion exchange, hydrophobic interaction or reversed-phase ligands may be coupled to such membranes ("affinity filters") to increase their selectivity for the target molecule by several orders of magnitude. Different membrane materials have been used for such purposes, including polyamide, regenerated cellulose, polystyrene, and various copolymers. Another advantage of membranes is their low flow resistance (backpressure). Since the separation efficiency (plate height) shows hardly any dependency on the flow rate, separations can be carried out within seconds at very high flow rates, unless the adsorption kinetics themselves pose a limit. Membrane chromatography is often performed using devices derived from conventional filtration units, although some filter holders exist, which are compatible to typical chromatographic systems or flow injection analyzers. In principle, five types of membrane units can be distinguished:

- Hollow fiber membranes (large dead volume causes zone broadening)
- Radial flow systems
- Dead end filter systems with single (or stack of) membrane(s)
- Compact porous disks
- Porous sheets loaded with specific binding particles (not strictly speaking membrane chromatography)

Membranes inserted in such a housing function as very short and wide chromatographic columns. The uniform distribution of the mobile phase over the relatively large cross-sectional area can be a problem, along with the collection of the target biomolecules without back-mixing and peak distortion. The scale up of membrane chromatography is usually straightforward, as ample experience exists in the scaling up of filtration units, and stacking of membranes of different functionalities can bring interesting possibilities in mixed-mode separations.

2. Monoliths

Chromatographic membranes were developed from typical filter membranes, which shows in their materials and also in the physical characteristics. Many have, for example, relatively broad pore size distributions. Lately the advantages of chromatographic membranes (low backpressure, superior mass transfer properties) have served

as impetus to design so-called high performance chromatographic disks (or monoliths), which are true high performance stationary phases for biochromatography. They typically consists of a rigid porous polymer (thickness millimeter range, diameter centimeter range) with a well-defined, narrow pore size distribution. The disks are prepared by *in situ* polymerization of suitable monomers and functionalized later. Examples of materials include rigid poly(glycidyl methacrylate-co-ethylene dimethylacrylate) (EDMA-GMA) networks, poly(styrene-co-divinylbenzene) copolymers, and silica phases obtained by a variant of the sol–gel technique.

SEE ALSO THE FOLLOWING ARTICLES

BIOMATERIALS, SYNTHESIS, FABRICATION, AND APPLICATIONS • BIOPOLYMERS • BIOREACTORS • ELECTROPHORESIS • INCLUSION (CLATHRATE) COMPOUNDS • LIQUID CHROMATOGRAPHY • MAMMALIAN CELL CULTURE • MASS TRANSFER AND DIFFUSION • MEMBRANES, SYNTHETIC, APPLICATIONS • PHARMACEUTICALS • PROTEIN SYNTHESIS

BIBLIOGRAPHY

Deutscher, M. P., ed. (1990). "Methods in Enzymology, Volume 182, Guide to Protein Purification," Academic Press, San Diego, CA.

Heftmann, E., ed. (1992). "Chromatography 5th Edition: Fundamentals and Applications of Chromatography and Related Differential Migration Methods," Journal of Chromatography Library, Volumes 51A–51B, Elsevier, Amsterdam.

Kastner, M., ed. (2000). "Protein Liquid Chromatography," Journal of Chromatography Library, Volume 61, Elsevier, Amsterdam.

Sofer, G., and Itagel, L. (1997). "Handbook of Process Chromatography: A Guide to Optimization, Scale-Up, and Validation," Academic Press, San Diego, CA.

Subramanian, G., ed. (1995). "Process Scale Liquid Chromatography," VCH, Weinheim.

Subramanian, G., ed. (1998). "Bioseparation and Bioprocessing—A Handbook," Vol. 1, Wiley—VCH, Weinheim.

Tissue Engineering

François Berthiaume
Martin L. Yarmush
Massachusetts General Hospital, Harvard Medical School, and Shriners Hospital for Children

I. A Brief History of Tissue Engineering
II. Fundamentals of Tissue Engineering
III. Applications of Tissue Engineering
IV. Future Prospects for Tissue Engineering

GLOSSARY

Allogeneic Qualifies tissues used for transplantation among different individuals of the same species.

Autologous Qualifies tissues used for transplantation to the same individual, or an identical twin, and thus not at risk of immune rejection.

Biomaterial A biocompatible material onto which cells can be cultured.

Bioreactor A device with special fittings which allows the large-scale culture of cells.

Connective tissue Tissue which primarily provides structural support in the body and is typically made of cells embedded in an extracellular matrix.

Convection A mode of transport driven by fluid flow carrying the solute particles to the cells or region where they are needed.

Differentiation The process whereby a cultured cell exhibits a greater number of characteristics reminiscent of the function and behavior of the parent tissue *in vivo*.

Diffusion A mode of transport driven by random molecular motion and which depends on the presence of a gradient of concentration of solute particles.

Endocrine cell A cell whose primary function in the body is to secrete factors which travel in the blood stream and regulate the function and metabolism of other cells.

Endothelial cell A cell which forms a selective barrier on the inner surface of blood vessels.

Epithelial cell A cell type which forms selective barriers that isolate different compartments from the rest of the body, such as the gut, stomach, or bladder.

Extracellular matrix Insoluble macromolecular network which surrounds cells in tissues.

Ligand Hormone or growth factor molecule which binds a specific receptor on the cell surface.

Morphogenesis The process whereby aggregates of cells undergo progressive reorganization into a tissue-like structure.

Receptor A specialized protein on the cell surface which binds specific hormones or growth factors and transmits this information inside the cell.

Signal transduction The process whereby the ligand–receptor binding event is intracellularly amplified and converted to a cellular response (i.e., such as cell division).

Stem cell An undifferentiated cell which has a high replicative potential and has the ability to convert into a wide variety of cell types expressing differentiated functions.

Xenogeneic Qualifies tissues used for transplantation across species.

TISSUE ENGINEERING can be defined as the application of scientific principles to the design, construction, modification, growth, and maintenance of living tissues. The main goals of tissue engineering are to help with the repair and regeneration of tissues *in vivo* and to grow tissues *in vitro* for use as models for physiological and pathophysiological studies, as well as to provide replacement parts for the body. Sometimes, tissues will repair and form scar tissue or tissue which does not exhibit a normal function and/or appearance. For example, tissue engineers have implanted polymeric tubes to promote the growth and reconnection of damaged nerves and used cultured skin grafts to cover deep burn wounds. Sometimes, the normal tissue regeneration process is too slow and temporary palliative care must be used to supply the vital missing functions to the patient. For example, tissue engineers are currently developing bioartificial liver assist devices for acute liver failure patients. Such devices may be used to buy time until a transplantable organ is available or may allow the patient's own liver function to return to recover, thereby obviating the need for liver transplantation altogether.

I. A BRIEF HISTORY OF TISSUE ENGINEERING

The use of materials derived from animal sources for making tissue replacement parts has been common practice for over 25 years with the use of bovine and porcine heart valves in cardiovascular surgery procedures. The source materials require special chemical processing and trimming before they are ready for use, all of which require extensive research and development. These devices may therefore arguably be considered the first tissue engineered devices used clinically. The function of these devices, however, is primarily a mechanical one and such implants do not significantly become repopulated with the host's cells (or, if they do, they do not significantly contribute to their function). More recently, biologically derived matrices, such as acellular human dermis (AlloDerm®, LifeCell, Inc.), which is made by treating human cadaver skin in such a way that no cells but the extracellular matrix remain, have been used in order to promote the regeneration of tissue in deep burn wounds.

The first man-made material designed to promote cell ingrowth and permanent incorporation into the body was developed by Ioannis V. Yannas (Massachusetts Institute of Technology) and John F. Burke (Massachusetts General Hospital and Shriners Burns Hospital, Boston) in 1980. It consists of a bovine collagen–glycosaminoglycan matrix made from chemical extracts of bovine skin and shark cartilage overlaid with a thin silicone membrane. This construct is applied onto deep burn wounds to faciltate the regeneration of the dermal layer of skin, after which the silicone sheeting is removed and replaced by a skin graft. This product was approved by the U.S. Food and Drug Administration (FDA) for clinical use in 1996 and is commercialized under the name of Integra® (manufactured by Integra LifeSciences, Inc.).

Advances in cell culture techniques have also played a pivotal role in the development of tissue engineered products. A landmark discovery by Howard Green and James Rheinwald (Harvard Medical School) in 1975 is the demonstration that keratinocytes from the skin epidermis could be cultured *in vitro* using a "feeder layer" of mouse fibroblasts. Keratinocytes were harvested from patients with extensive burns and propagated *in vitro* until the available surface area of cultured skin was sufficient to use as an autologous grafting material. In 1988, Genzyme Corp. began the Epicel® service and more recently extended this concept to the propagation of chondrocytes for the treatment of cartilage defects in knee joints (Carticel®).

Other pioneering work in tissue engineering includes studies published by Eugene Bell (Massachusetts Institute of Technology) between 1979 and 1981 describing the first matrix–cell composite grown *in vitro* prior to *in vivo* implantation. Human dermal fibroblasts were seeded into collagen gels to produce a bioartificial dermis, which was then overlayed with a monolayer of epidermal cells (keratinocytes) to generate a full-thickness skin equivalent. This product, available under the name of Apligraf® (Organogenesis, Inc.), was approved by the FDA in 1998 for treating nonhealing venous ulcers and, more recently, diabetic foot ulcers. Unlike autologous skin grafts, which require several weeks to become available after harvesting the source cells from the patient, Apligraf® is made of allogeneic cells obtained from donated human foreskins and is a ready- to-use product available within a short notice. On the other hand, since Apligraf® contains allogeneic cells, it eventually becomes rejected by the recipient's immune system, and repeated treatments may be necessary until the ulcer heals on its own.

Beyond the few tissue engineered products which are currently used clinically, there are many others in the pipeline of biotechnology companies as well as research laboratories around the world. One important area of tissue

TABLE I Companies Selling Tissue Engineered Products and Products for Regenerative Medicine

Company	Product name(s)	Application(s)	Core technology
Edwards Lifesciences; Irvine, CA	Carpentier-Edwards PERIMOUNT pericardial valve, Edward Prima Plus Stentless Bioprosthesis	Heart valve replacements	Chemically treated xenogeneic heart valve tissue
Medtronics; Minneapolis, MN	Hancock® II aortic and mitral bioprostheses	Heart valve replacements	Chemically treated xenogeneic heart valve tissue
St. Jude Medical; St. Paul, MN	Toronto SPV® valve	Heart valve replacements	Chemically treated xenogeneic heart valve tissue
LifeCell; Branchburg, NJ	AlloDerm® acellular human dermis	Burn wounds	Acellular dermis from human cadavers
Regeneration Technologies; Alachua, FL	Regenapack™ regeneration template, CorlS™ and AlloAnchor™ bone-healing screws and pins	Bone wound healing	Precision-tooled natural bone matrices
Integra LifeSciences; Plainsboro, NJ	Integra® dermal regeneration template, BioMend® absorbable collagen membrane	Burn wounds, periodontal disease	Collagen-based matrices
Sulzer Medica; Winterthur, Switzerland	Ne-Osteo™ osteogenic bone-filling material	Spinal fusion, periodontal disease	Collagen-based matrices with growth factors
Curis; Cambridge, MA	OP-1 Implant™ osteogenic bone-filling material	Nonunion fractures	Collagen-based matrices with growth factors
Genzyme; Cambridge, MA	Carticel® autologous chondrocytes, Epicel® autologous keratinocytes	Cartilage defects of the knee, burn wounds	Culture of autologous keratinocytes and chondrocytes
Advanced Tissue Sciences; La Jolla, CA	TransCyte™, Dermagraft®	Burn wounds and skin ulcers	Bioreactors for three-dimensional stromal cell culture technologies
Organogenesis; Canton, MA	Apligraf®	Skin ulcers	Dermal-epidermal composites
Ortec International; New York, NY	Composite cultured skin (CSS)	Epidermolysis bullosa, skin ulcers	Dermal-epidermal composites
Circe Biomedical; Lexington, MA	HepatAssist® liver support device	Extracorporeal liver-failure treatment	Hepatocyte bioreactors
Vitagen; La Jolla, CA	ELAD™ liver support device	Extracorporeal liver-failure treatment	Cultured human hepatoma cell line
University of Berlin; Hybrid Organ GmbH, Berlin, Germany	Modular Extracorporeal Liver Support System (MELS)	Extracorporeal liver-failure treatment	Hepatocyte bioreactors

engineering still in the early stages is the development of bioartificial organs such as pancreatic islets and livers, some of which consist of complex devices that incorporate large cell numbers into novel bioreactor systems. Table I provides a more exhaustive listing of companies currently involved in the production of tissue engineered products approved for clinical use or which are in the more advanced stages of clinical trials.

The early discoveries and major advances in tissue engineering have been in large part the result of empirical studies because of our lack of understanding of the basic phenomena that control tissue formation and repair. Tissue engineering is now also evolving as a science which uses the theoretical framework of core engineering disciplines, including thermodynamics, transport, reaction kinetics, and control theory. The basic understanding of the rules that govern tissue repair, regeneration, and development will enable one to predict the behavior and performance of more complex tissue engineered constructs, a necessary step towards the optimization of tissue engineered products. The first such studies were published by Malcom S. Steinberg (University of Princeton) in the early 1960s and describe rules governing cell-sorting phenomena in multicellular systems containing different

cell types. Although the initial motivation for these studies was to undertand the mechanisms of embryonic development, the derivations are also relevant to the engineering of tissues for clinical applications. In the 1970s, several studies by Douglas A. Lauffenburger (then at the University of Illinois, now at M.I.T.) and Robert T. Tranquillo (University of Minnesota) set the stage for the modeling of intracellular signaling processes as well as cell-migration phenomena.

II. FUNDAMENTALS OF TISSUE ENGINEERING

A. Biomaterial Design

1. Materials Used in Tissue Engineering

The vast majority of mammalian cells are anchorage dependent and therefore must attach and spread onto a substrate to proliferate and function normally. While in traditional tissue culture systems two-dimensional surfaces are used to grow cells, tissue engineering often requires the use of three-dimensional matrices which allow cell ingrowth and organization reminiscent of actual tissues found *in vivo*. The choice of extracellular matrix material is highly dependent on the intended use of the tissue (whether its function is structural or biochemical or both) and on the respective roles of materials and cells in the reconstructed tissue. A list of three-dimensional materials used in tissue engineering is given in Table II.

Matrices derived of naturally occuring tissues, such as animal-derived heart valves, acellular dermis, and bone-derived matrices, are typically of allogeneic and xenogeneic origin. They are prepared via physical and chemical treatments, such as freeze-drying, cross-linking by glutaraldehyde, and detergent-mediated removal of cells, in order to enhance their physical properties and remove any antigen-bearing cells which could trigger undesirable immune responses. These materials retain the chemical composition and microarchitecture proper to the tissue that they are derived from, which can enhance their function. For example, blood vessel growth into an acellular dermis applied onto a burn wound will preferentially occur in the spaces formerly occupied by the blood vessels in the original intact tissue. Soluble factors are often retained within the matrix and can have pro-angiogenic (small intestinal mucosa) or anti-angiogenic (amniotic membrane) properties. A disadvantage of these materials is that their chemical composition is often only partially known, availability may be limited, and issues such as batch-to-batch variation and potential contamination with pathogens must be addressed on a continuous basis.

Most of the natural extracellular matrix materials, except bone, can be at least partially solubilized by chemical processing and reconstituted into three-dimensional gels of any shape or form. Although the microarchitecture is lost, these reconstituted matrices retain many chemical features of the extracellular matrix proteins including bound growth factors found in the original material. Commonly used reconstituted matrices include type I collagen

TABLE II Materials Commonly Used in Tissue Engineering

Name	Composition	Applications
Intact extracellular matrices		
Amniotic membrane	Collagen, fibronectin, laminin, GAG, growth factors	Corneal epithelium
Acellular dermis	Collagen, laminin, elastin	Skin epithelium
Small intestinal mucosa	Collagen, fibronectin, GAG, growth factors	Smooth muscle (vascular, urogenital)
Carbonate apatite (dahllite)	Calcium/magnesium carbonate/phosphate	Bone
Reconstituted extracellular matrices		
Type I collagen gel	Collagen	Skin dermis, tendon, hepatocyte
Collagen–GAG[a] complexes	Collagen, GAG	Skin dermis, tendon, nerve guidance
Engelbreth–Holm–Swarm tumor matrix gel (Matrigel)	Collagen, laminin, GAG, growth factors	Hepatocyte
Synthetic matrices		
Carbonate apatite	Calcium/magnesium carbonate/phosphate	Bone
pLA/pLGA co-polymer	Poly(lactic-*co*-glycolic) acid	Cartilage, bone, epithelium (gut, urogenital), hepatocyte
Dacron®	Polyethylene teraphtalate	Vascular endothelium
Gore-Tex®	Expanded polytetrafluoroethylene	Vascular endothelium
pHEMA/MMA co-polymer	Poly(hydroxyethyl methacrylate)	Vascular endothelium

[a] GAG = glycosaminoglycan.

which is isolated from rat tail or bovine skin by mild acid treatment. The acid solution of collagen can be induced to form a gel upon restoring a physiological pH of 7.4, which causes the polymerization of collagen molecules into a large network of fibrils. The extent of cross-linking in this collagen is very low in comparison with that of the native tissue, and as a result reconstituted collagen gels undergo rapid proteolytic degradation *in vivo*. To remediate this problem, chemical cross-linking is induced by either glutaraldehyde or dehydrothermal (vacuum and $\sim$100°C) treatment. For example, the skin substitute Integra® is made of a mixture of solubilized collagen and glycosaminoglycans whose extent of cross-linking has been optimized to withstand the specific environment of dermal wounds such as nonhealing ulcers and deep burns.

Materials used in tissue engineering must be able to withstand physical forces to which they are subjected. These forces naturally occur in load-bearing tissues such as bone and cartilage, as well as in other applications such as blood vessels, which must have burst pressures exceeding arterial levels. Physical forces can also be generated by the cells making up the bioartificial tissue, as cells have been shown to exert tractional forces on their points of attachment. Known examples of the effect of cell tractional forces include the contraction of collagen gels by fibroblasts and the formation of "ripples" by cells placed on thin flexible silicone sheets. Specific mechanical properties are required in certain applications, especially in the case of artificial vascular grafts, which must exhibit the same compliance as that of normal blood vessels. The mismatch in compliance that often occurs between the host's vessel and the graft is believed to be be an important factor leading to artificial vascular graft failure *in vivo*.

Systems using cells that do not secrete a structurally dense extracellular matrix must rely on the synthetic matrix provided to retain their structural integrity. The matrix must be able to withstand both the weight of the cultured cells as well as tensile forces generated by cells growing on the substrate. The use of relatively fluid substrates induces different cellular morphologies than does the use of rigid surfaces. Because fluid substrates cannot oppose cell-generated forces, cell–cell adhesive forces predominate over cell–substrate adhesive forces, which leads to cell aggregation as seen with hepatocytes plated on heat-denatured collagen as opposed to type I collagen. In high-density, three-dimensional cultures, cell-generated forces may become significant as seen with fibroblasts that can dramatically reduce the volume of collagen lattices.

Naturally derived matrices provide good substrates for cell adhesion because cells express the adhesion receptors which specifically recognize and bind to extracellular matrix molecules which make up these matrices. Nevertheless, there have been considerable advances in the development of synthetic biocompatible polymers, which theoretically have an unparalleled range of physical and chemical properties. In practice, however, most tissue engineering development has been limited to using a relatively small number of man-made materials, in part due to a reluctance to expend time and money to secure regulatory approval for clinical use of untested biomaterials. The most extensively used materials in medicine are titanium and inert plastics such as Teflon® for orthopedic applications and artificial vascular graft prostheses, respectively. Typical problems encountered with artificial orthopedic materials include failure of the graft–host tissue interface in the case of bone substitutes, which may be due to an adverse reaction to the artificial material, and progressive wear-and-tear in artificial joints, which do not have the ability to regenerate and repair, unlike natural joint surfaces. Artificial vascular grafts tend to activate the blood clotting cascade and may also cause thickening of the vascular tissue near the point of attachments to the host's vascular tree. These responses do not pose a major problem for the function of large-diameter grafts (e.g., thoracic aorta), but have prevented their use as smaller vessels such as coronary bypass segments, for which the demand is very high. Finally, all artificial materials implanted *in vivo* are highly susceptible to colonization by bacteria which can form biofilms highly resistant to antibiotics. Furthermore, recent studies suggest that the function of immune cells may also be compromised on certain artificial surfaces, which reduces the ability of the host to clear infections.

To overcome the problems due to foreign-body reactions caused by artificial materials, there is currently heightened interest in the use of biocompatible polymers which naturally degrade *in vivo*. One of the best known and most commonly used synthetic biodegradable polymers in tissue engineering are the poly(lactic-*co*-glycolic) acid copolymers, which have been used in the form of biodegradable sutures for several decades. In 1988, Robert S. Langer (Massachusetts Institute of Technology) and Joseph P. Vacanti (Children's Hospital, Boston) pioneered their use in tissue engineering. Currently used as part of skin substitutes commercialized by Advanced Tissue Sciences, they are now the most widely investigated artificial biodegradable polymers in tissue engineering, with applications including cartilage, bone, and various epithelia (intestine, bladder, liver). This material hydrolyzes completely within weeks, months, or years, depending on the exact composition (in general, increased hydrophobicity correlates with a decreased degradation rate) and thickness. Cells migrating in from surrounding tissues after implantation *in vivo*, or cells directly seeded into the polymer, secrete their own extracellular matrix which gradually replaces the polymer scaffold as the latter slowly dissolves away. It is important that the degradation rate of

biodegradable extracellular matrix materials, when used deliberately, must be such that the cell-generated matrix has sufficient time to form and the mechanical integrity of the tissue is maintained at all times. In the end, there are no foreign materials left in the patient to cause adverse long-term immune reactions or harbor bacterial infections.

2. Optimization of Surface Chemistry

Cells do not usually directly attach to artificial substrates, but rather to extracellular matrix proteins which are physically adsorbed (i.e., by virtue of hydrophobic and electrostatic interactions) or chemically attached (i.e., via covalent bonds) to the surface. Many polymers are highly hydrophobic and do not favor protein adsorption. Increasing substrate wettability to a certain point (water contact angle of 60° to 80°), such as by using ionized gas, increases protein adsorption and is commonly used for preparing tissue-culture-grade polystyrene Petri dishes. This process only modifies the surface of the material, and thus minimally affects its bulk mechanical properties. Highly hydrophilic surfaces are also not favorable to protein adsorption. In addition, if negatively charged, they may cause repulsive electrostatic interactions with the cells, the latter of which usually display a negative surface charge due to the presence of negative sialic acid residues on their surface glycocalyx. Conversely, coating surfaces with positively charged materials, such as poly-L-lysine, has been used to promote cell adhesion to the surface. Physisorption of proteins for which cells do not express any adhesion receptor, such as albumin, is also commonly used to prevent cell adhesion.

Physisorbed proteins are not stably bound and can be displaced by other proteins. This especially occurs in complex media such as plasma, where fibrinogen physisorption may occur within seconds, following by displacement of more slowly diffusible but "stickier" proteins (this is sometimes called the *Vroman effect*). When the surface is transferred to a different medium after protein coating, the type and amount of physisorbed proteins will change until reaching equilibrium with the proteins in solution above the surface. The time scale for desorption can extend over several hours, thus physisorption can be useful to control the initial attachment of cells at the time of seeding. Over a period of several days of culture, however, virtually all cells will have secreted significant quantities of their own extracellular matrix onto the substrate, and the initial surface properties of the material often become irrelevant.

Because physisorption is notoriously nonselective, covalent modification of substrates or chemisorption is used if it is necessary to provide more control over the type, density, and distribution of adhesive protein on the surface of the material. For this purpose, several chemical processes are available depending on the type of surface to be modified (Table III). The first step involves using a reactive chemical that bonds to the surface and has a free functional group that easily reacts with free thiol, hydroxyl, carboxyl, or amine groups on proteins. This step often requires harsh chemical conditions, while the second step, which involves conjugation of the protein, can be done under physiological conditions. This approach is also suitable to graft small adhesive peptides (e.g., RGD) which otherwise would not stably bind to surfaces by physisorption (Table IV). Furthermore, physisorption sometimes leads to unexpected changes in protein activity, probably due to denaturation on the surface. For example, adsorbed fibrinogen activates and binds to platelets, unlike solution-phase fibrinogen in normal plasma or blood.

TABLE III Ligands Used for Chemisorption of Protein to Surfaces

Ligand	Type of primary surface
Silanating reagents	Glass, silicon
Alkane thiols	Gold
Carboxylic acids	Alumina
Sulfonyl halides, carbonyldiimidazole, succinimidyl chloroformate, succimidyl esters	Synthetic polymers[a]

[a] Dacron and PTFE require chemical treatment in order to create free alcohol and carboxylic groups prior to derivatization.

TABLE IV Compounds Used to Promote or Prevent Cell Adhesion to Surfaces

Pro-adhesive	Anti-adhesive
Extracellular matrix proteins	Polyethylene glycol
Collagen	Albumin
Fibronectin	Polyvinyl alcohol
Vitronectin	Cellulose acetate
Laminin	Agarose
	Sulfonate residues
Adhesive peptide sequences[a]	
RGD (from collagen)	
YIGSR, IKVAV (from laminin)	
REDV (endothelial-specific)	
Adhesion molecules	
Intercellular adhesion molecule-1 (ICAM-1)	
Vascular cell adhesion molecule-1 (VCAM–1)	
Platelet cell adhesion molecule-1 (PCAM-1)	
Sialyl Lewis X	

[a] Single-letter amino acid abbreviations.

The design of more sophisticated cultured tissues using more than one cell type can be enhanced by spatially controlling the seeding process. For this purpose, various methods for patterning the deposition of extracellular matrix or other cell attachment factors onto surfaces have been developed. Photolithography involves spin-coating a surface (typically silicon or glass) with an $\sim$1-μm thick layer of photoresist material, exposing the coated material to ultraviolet light through a mask containing the pattern of interest, and treating the surface with a developer solution which dissolves the exposed regions of photoresist only (Fig. 1). This process leaves photoresist in previously unexposed areas of the substrate. The exposed areas of substrate can be chemically modified for attaching proteins, etc., or can be treated with hydrofluoric acid to etch the material. The etching time controls the depth of the channels created. Subsequently, the leftover photoresist is removed using an appropriate solvent, which leaves a surface patterned with different molecules and/or grooves. A disadvantage of this method is that it uses chemicals toxic to cells and generally harsh conditions which could denature proteins are used.

The etched surfaces produced by photolitography can be used to micromold various shapes in a polymer called poly(dimethylsiloxane) (PDMS). The PDMS cast faithfully reproduces the shape of the silicon or glass mold to the micron scale and can be used in various "soft lithography" techniques, including microstamping, microfluidic patterning, and stencil patterning (Fig. 2). An infinite number of identical PDMS casts can be generated from a single master mold, which makes the technique very inexpensive. Soft lithography methods can be used on virtually any type of surface, including curved surfaces, owing to the flexibility of PDMS. Another patterning method which works well at larger size scales is microprinting using laserjet technology, which can also be used to create three-dimensional structures (Fig. 3).

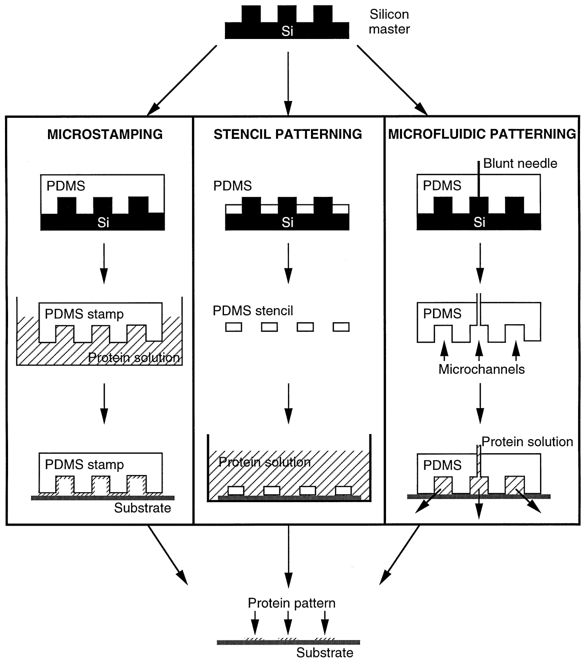


FIGURE 2 Patterning using soft lithography. The silicon master is used as a mold to create a flexible replica made of poly(dimethylsulfoxane) (PDMS). The replica can be used as a stamp to deposit protein on a substrate, as a stencil to cover up selected regions of the substrate during protein coating, or as a series of flow channels to deliver a protein-coating solution onto the substrate.

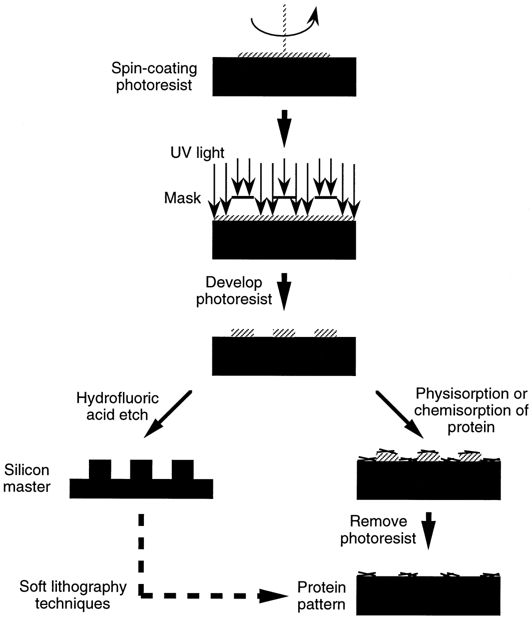


FIGURE 1 Patterning using photolithography. A silicon (or glass) wafer coated with photoresist is exposed to ultraviolet light in areas determined by a mask overlay. A developer chemical selectively removes the photoresist and the exposed areas of silicon can be either etched for use in soft lithography techniques (see Fig. 2) or coated with proteins.

In using these approaches, it is important that the base material be resistant to physisorption, or the selectivity of the adhesive groups may be significantly reduced *in vivo*. A successful approach to prevent adhesion to the base material is via covalent attachment of anti-adhesive factors on the remaining functional groups.

Micropatterning is especially desirable to maximize heterotypic cell–cell interactions between a parenchymal cell such as hepatocyte and supporting or "feeder" cells such as fibroblasts. Keeping in mind that cells cultured on surfaces do not usually layer onto each other (except for malignant cancer cell lines), random seeding using a low ratio of parenchymal cells to feeder cells will achieve this goal, but at the expense of using a lot of the available

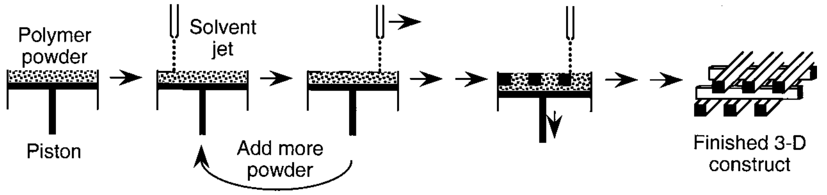


FIGURE 3 Microprinting three-dimensional scaffolds. A small jet of solvent is sprayed onto a packed bed of polymer powder to induce bonding of the powder into a solid in selected regions. Alternating solvent spraying and new additions of polymer powder eventually creates a three-dimensional shape suitable for cell culture.

surface for fibroblasts, which do not provide the desired metabolic activity. On the other hand, micropatterning techniques enable optimization of the seeding pattern of both cell types so as to ensure that each hepatocyte is near a feeder cell while minimizing the number of feeder cells. As a result, metabolic function per area of culture is increased and the ultimate size of bioreactor with the required functional capacity is reduced.

3. Fabrication of Porous Matrices

Porous matrices are often used to reconstruct connective tissues because they allow the formation of complex extracellular matrix networks responsible for the tissue's mechanical properties and the fusion of the implant with the host's tissue. Pore sizes in the range of 30 to 300 μm are the most common. Smaller pore sizes provide more surface area per volume of matrix; however, pores less than 30 μm will not allow seeding or ingrowth of the host's tissue into the matrix.

Porous materials are usually prepared by salt-leaching or freeze-drying techniques. The first method involves adding water-soluble crystals (e.g., NaCl) of size range similar to the desired pores to the melted base polymer material. After solidification of the polymer, the salt crystals in the resulting solid are dissolved by exposure to aqueous solutions, leaving a pore in the place of every crystal. An alternative approach is the use of supercritical carbon dioxide to create pores by induction of microbubble formation within the polymer.

The freeze-drying technique is based on the general principle that when freezing a solution, the solvent forms pure solid crystals while all solute materials are concentrated in the remaining unfrozen fraction. During the subsequent drying process, the solid crystals evaporate and leave pores. The morphology of the solid crystals is dependent on the physico-chemical properties of the solution, the temperature gradient at the liquid–solid interface, and the velocity of that interface. The directional solidification system shown in Fig. 4 allows one to independently control each one of these three parameters. During directional solidification, the size and shape of the crystals forming can be predicted from basic physics principles.

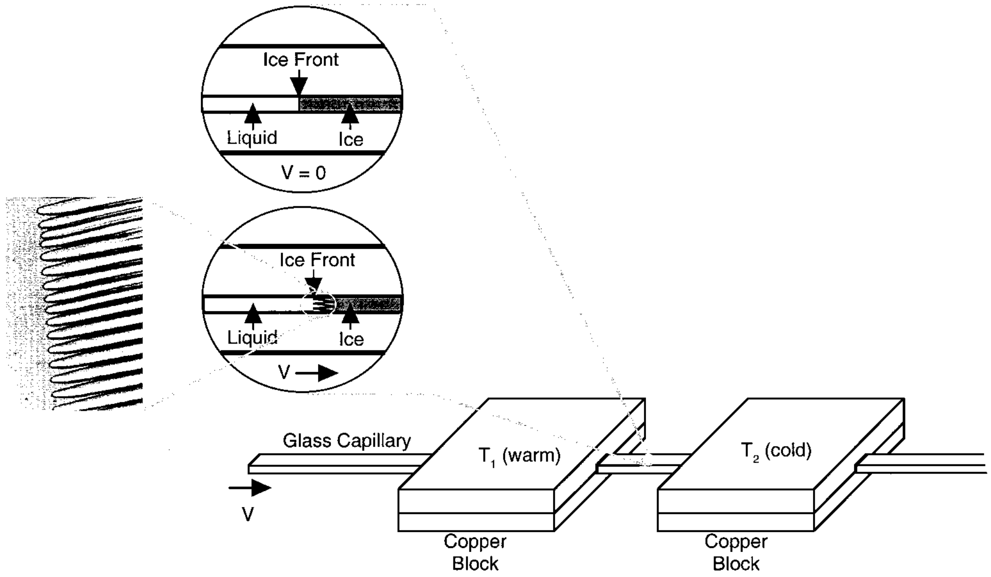


FIGURE 4 Directional solidification stage to pattern crystal formation during freezing of polymer solutions. Inset on left shows the morphology of water crystals during freezing of a collagen solution in isotonic saline with 1 m*M* HCl.

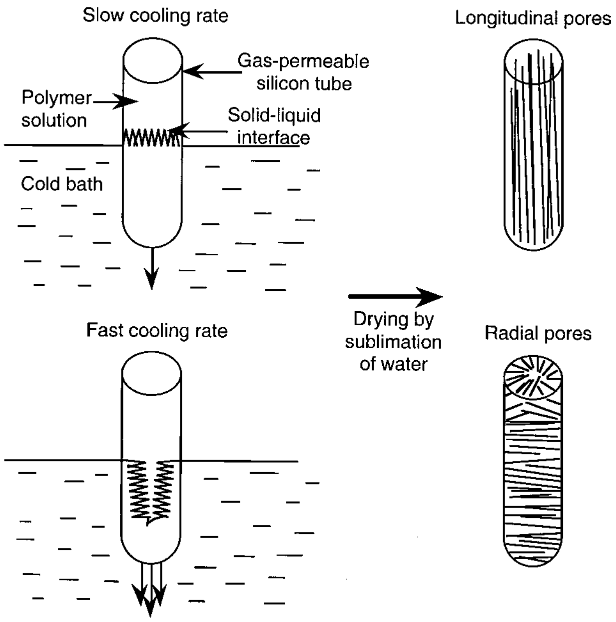


FIGURE 5 Method for creating oriented pores in cylindrical matrices. Slow cooling (top panel) promotes solidification from the solid–liquid interface, thus generating vertically oriented crystals. Fast cooling (bottom panel) promotes solidification from the walls of the tube, leading to horizontally oriented crystals. After drying, the orientation of the pores reflects that of the crystals.

Furthermore, the solid crystals tend to orient in the direction of the temperature gradient, so that the direction of the pores can be controlled as well. For practical applications, however, it is more typical to freeze solutions containing biomaterials in a bulk fashion. For example, nerve guidance tubes have been produced by immersion of suspensions of collagen–GAG complexes contained in gas-permeable silicone tubes in a cold bath. As depicted in Fig. 5, a slow rate of immersion causes the formation of crystals (and pores, eventually) predominantly oriented along the length of the tube, which is the geometry desired for this application. More rapid immersion, on the other hand, would lead to crystal growth primarily in the radial direction. It is noteworthy that the rate of freezing and the temperature gradient are difficult to control and maintain constant throughout the freezing process. Thus, typically, porous materials made by this technique exhibit nonuniform pore sizes as one moves from the surface to the center.

B. Cell Engineering

1. Growth Factors, Hormones, and Signal Transduction

Cultured cells are often included in polymer scaffolds used in tissue engineering to make up for the limited potential of the host's surrounding cells to regenerate the damaged or missing tissue. Control of the function and growth of the cells is critically important and may require the use of exogenous growth factors, which are small proteins that act as ligands binding to specific cognate receptors on target cells. Table V provides a sample list of growth factors currently used in tissue engineering. Hormones are smaller compounds which have similar effects and include small peptides as well as a range of lipid-soluble compounds derived from cholesterol and fatty acids. Hormones and growth factors may be incorporated into the scaffold itself and released over time, or the cells in the construct can be modified genetically (see section on genetic engineering) or otherwise to produce the growth factors themselves. In addition, the recipient's own tissue surrounding a tissue engineered implant may undergo an inflammatory response due to the surgical trauma or the presence of impurities (e.g., bacterial-derived lipids such as endotoxin), as well as immunogenic factors, including proteins of animal origin. There are several soluble mediators released during the course of an inflammatory response, some of which can either stimulate or suppress cell growth as well as other cellular functions in the implant.

TABLE V Growth Factors Commonly Used in Tissue Engineering

Growth factor	Target cell
Epidermal growth factor	Keratinocytes, hepatocytes
Hepatocyte growth factor, scatter factor	Epithelial cells
Interleukin-2	White blood cells
Platelet-derived growth factor (PDGF)	Fibroblasts, smooth muscle cells
Fibroblast growth factor	Fibroblasts, smooth muscle cells, endothelial cells
Vascular endothelial growth factor (VEGF)	Endothelial cells
Nerve growth factor	Neurons
Insulin-like growth factor	Muscle, keratinocytes
Osteogenic protein 1	Osteoblasts

In order to analyze, predict, and optimize the cellular response to growth factors, mathematical models can provide useful insights. The system which has been the most extensively studied with respect to the quantitative aspects of receptor-mediated signaling is that involving epidermal growth factor (EGF) binding to its receptor (Fig. 6). A feature of the early signaling events is the internalization of the EGF–EGF receptor complexes into the cell, which can then be recycled back to the cell surface or degraded within the cell. First-order kinetic mass balance equations can be derived for the species shown in Fig. 6. The binding of growth factors to receptors and ensuing

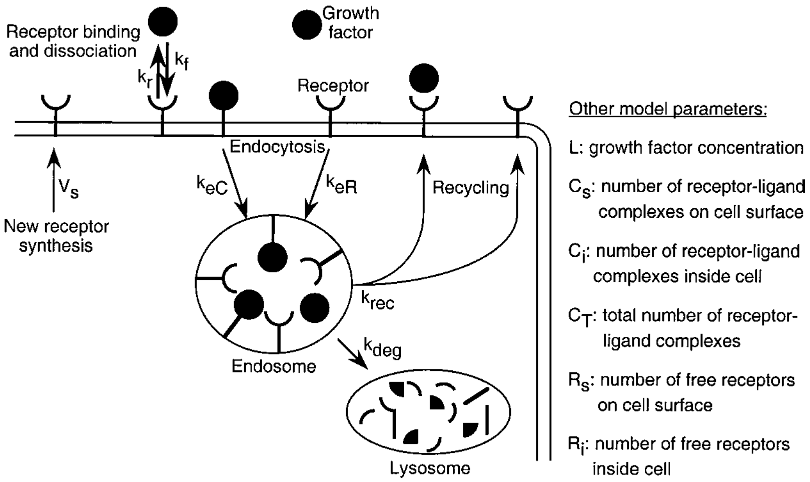


FIGURE 6 Simplified model for the binding and fate of a ligand (growth factor) binding to its receptor on the cell surface. After binding, the ligand–receptor complexes are internalized into an endosomal compartment. Its contents can then be recycled to the cell surface or fuse with a lysosome and undergo degradation. The *k* parameters shown represent first-order rate constants for each process shown.

intracellular changes occur in the time scale of seconds to minutes, and a steady state is reached within one hour. In many cases, it can be assumed that the cellular response is proportional to the total number of receptor-bound growth factor molecules. Since in most cases growth factor levels also do not change very quickly in the environment, one can assume a pseudo steady state and obtain the following results for the number of EGF–EGF receptor complexes:

$$C_s = \left(\frac{K_{ss}L}{1+K_{ss}L}\right)\frac{V_s}{k_{eC}}, \quad K_{ss} = \frac{k_{eC}k_f}{k_{eR}(k_{rec}+k_{eC})} \tag{1}$$

$$C_i = \left(\frac{k_{eC}}{k_{\text{deg}}}\right)C_s \tag{2}$$

$$C_T = C_s + C_i \tag{3}$$

$$R_s = (k_r + k_{eC})\frac{C_s}{k_f L} \tag{4}$$

Using accepted parameter values for this system, one can notice that the proportion of complexes that are intracellular increases with EGF concentration (Fig. 7). Furthermore, the total number of bound receptors in the cell as well as on the cell surface reaches a maximum corresponding to about 25% of the total number of receptors. The process of receptor-mediated endocytosis thus limits the number of receptors available for binding at any time, which is a well-known mechanism of downregulation of growth factor responses. Several studies have shown that a more sensitive response to a particular growth factor may be obtained using modified growth factors or cells with altered receptors which reduce the rate of internalization of the complexes. Another strategy to reduce the rate of internalization is to immobilize the growth factor to the surface of the substrate to which the cells are attached. This has been shown to work in the case of EGF and fibroblasts; however, in other systems, the activity of the growth factor may be partially or completely lost. The EGF–EGF receptor model can be further refined by taking into account the removal of growth factor from the extracellular medium, the effect of receptor clustering, several pathways operating at different rates, etc. Cells that produce their own growth factors as part of an autocrine loop can also be modeled in a similar fashion by adding appropriate terms for growth factor release and diffusion around the cells. Such a model may be useful to predict the effect of cell density on growth rate as well as other density-dependent functions.

A difficult problem which remains in this area is to relate, on a theoretical basis, the receptor–ligand binding phenomena on the cell surface to the observed cellular response. The intracellular signaling pathways typically function as a cascade of events leading to the sequential activation of intracellular signaling molecules, often by phosphorylation of specific amino acid residues on the signaling proteins. One of the final targets of this cascade may be one or several transcription factors, specialized proteins that have the ability to migrate into the cell nucleus and trigger the synthesis of new proteins or alterations in cell behavior, such as cell division. A large number of growth factors trigger the mitogen-activated protein (MAP) kinase cascade (Fig. 8), and there is evidence that similar cascades exist for other mediators. Kinetic modeling of each step as a reaction, using parameters determined through analyses in cellular extracts of the intracellular concentrations and kinetic properties of

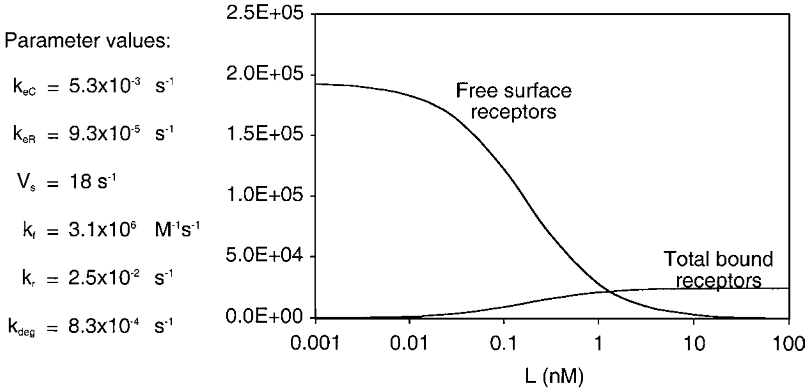


FIGURE 7 Effect of ligand concentration on the number of remaining free receptors on the cell surface and the total number of bound receptors. Parameter values used (shown on the left of graph) are literature values for epidermal growth factor and fibroblasts.

individual components of the cascade, reveals that the cascade operates as a signal amplification system that tends to produce a switch-like behavior in the cellular response to growth factors. Thus, the response of a cell to growth factors can often be represented by a threshold model where no effect occurs below that threshold and a maximal response occurs above that threshold. When studying a cell population, however, a dose-dependent response may still be observed because cells tend to exhibit a distribution in the concentration of each effector molecule involved in the signaling cascade. It is important to note that modeling of cellular responses is still in its infancy, as there are no models yet published addressing the effect of multiple growth factors and multiple signaling cascades operating at the same time, which would be more typical of conditions used to culture cells in tissue engineering applications.

Besides soluble growth factors and hormones in the medium, there are many other parameters in the environment that influence cell growth rate and the expression of specific functions, which may be desirable or even necessary for tissue engineering applications. For example, controlling the density of adhesion sites is a potentially powerful means of controlling cell growth and function. The spreading of a cell on a surface increases together

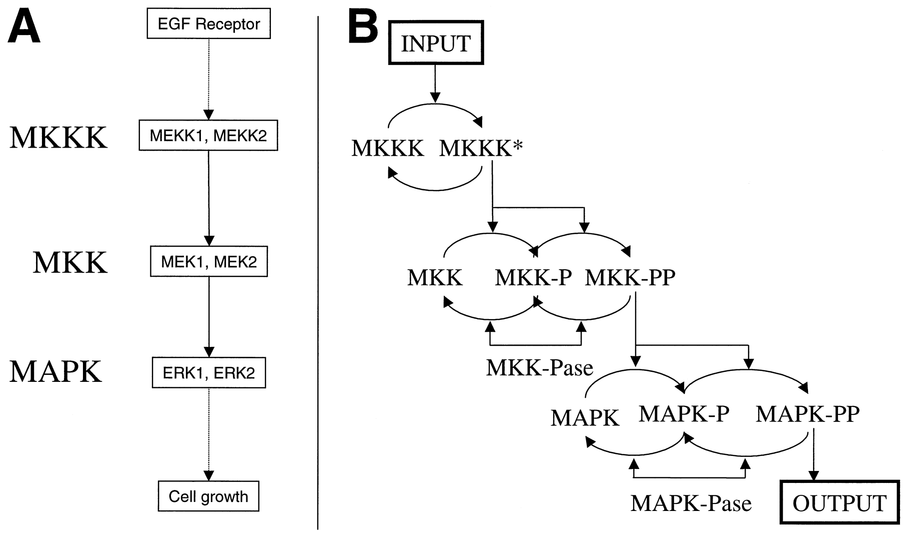


FIGURE 8 The signaling cascade triggered by a typical growth factor. MAPK = mitogen-activated protein kinase; MKK = MAPK kinase; MKKK = MKK kinase. MKKK is activated by receptor–ligand binding. Activated MKKK then phosphorylates MKK twice sequentially to activate MKK. Activated MKK then phosphorylates MAPK. MAPK activates factors (not shown) which migrate into the cell nucleus, bind the cell's DNA, and initiate cell replication. The multistep nature of the cascade causes amplification of the signal generated by the receptor–ligand binding event to trigger a switch-like cellular response. MEKK, MEK, and ERK are the names of kinases specific to the epidermal growth factor pathway.

with the surface density of extracellular matrix molecules, which is typically accompanied by an elevation in DNA synthesis and proliferation rates.

A general rule of thumb in cell culture techniques is that proliferation and differentiation are mutually exclusive. In other words, conditions promoting the expression of differentiated functions are often not optimal for replicating cells. For example, fibroblast growth factor-stimulated capillary endothelial cells plated on nonadhesive surfaces coated with decreasing concentrations of fibronectin switch from a spreading to a tubular capillary-like shape, with a concomitant reduction in cell growth. In some cases, cell differentiation can also be induced by altering the culture environment so as to mimic a subset of *in vivo* conditions. For example, keratinocytes, a type of epithelial cell which forms the epidermal component of the skin, can be propagated *in vitro* using a serum-free culture medium; a single human neonatal foreskin can provide enough cells to yield over 100 m^2 of graftable tissue. Cells in cultured epidermal sheets are not well differentiated but exposure to air while in culture or after grafting onto the host induces the formation of a stratified differentiated epidermis.

One of the challenges of tissue engineering is to produce large cell masses that are well differentiated. Although differentiated cells do not always proliferate easily *in vitro*, it may be possible to optimize culture conditions to stimulate cell propagation and then to change these conditions so that a stable and functional phenotype is exhibited by the cells. For example, chondrocytes seeded on plastic in the presence of serum proliferate but secrete a significant amount of type I collagen and small proteoglycans, which are not normally found in cartilage. Embedding these cells in an agarose gel induces the re-expression of the normal phenotype found *in vivo*, which is characterized by the production of type II collagen and deposition of large aggregating proteoglycans.

2. Genetic Engineering

While control of the extracellular environment remains the primary means of modulating cell function and proliferation in tissue engineering, it is sometimes advantageous to alter the genetic make-up of cells to extend their basic capacity to perform specific functions. Describing the techniques used for genetically altering cells is beyond the scope of this chapter, and the reader is referred to the numerous texbooks and reviews on the subject. Genetic modification of cells in tissue engineering has included the following applications: (1) expressing functions not normally present in a particular cell type or overexpressing existing functions, and (2) expressing "immortalizing" genes or genes that protect cells against death caused by apoptosis. The first application may be part of a gene therapy protocol aimed at providing a patient who has a single enzyme deficiency (e.g., adenosine deaminase) with implantable cells to perform the missing function. Another important application is the (over)expression of angiogenic factors that promote the rapid invasion by blood vessels and vascularization of implantable tissue constructs. Immortalizing genes, such as the viral SV40 T antigen and telomerase are primarily used to promote the replication of cells typically very difficult to grow *in vitro*, such as hepatocytes, pancreatic beta cells, etc. The use of anti-apoptotic genes in tissue engineering is a relatively new trend stimulated by the difficulties of maintaining cell viability in large tissue constructs made of cells sensitive to the depletion of nutrients—for example, in the case of hepatocytes in bioartificial livers.

One of the issues raised by the use of genetic engineering in tissue engineered products is the unknown effects of persistent expression of the transgene in the implanted cells. For example, overexpressing growth factors may be beneficial to the process of growth and integration of a engineered tissue implanted in a host; however, the long-term effects of high levels of growth factors are unknown and could perhaps be detrimental. This problem may be resolved soon, however, with the advent of new molecular biology techniques that allow for the "excision" at will of the transgenes in order to restore the native state of the cells.

3. Metabolic Engineering

Metabolic engineering has been defined as the introduction of specific modifications to metabolic networks for the purpose of improving cellular properties. In recent years, metabolic engineering has gained importance in biotechnology, being used largely to improve existing processes involving the production of chemicals using microorganisms. Although less widely appreciated, metabolic engineering techniques can be applied to study physiological systems and isolated whole organs *in vivo* to elucidate the metabolic patterns that occur in different physiological states, such as fed, fasted, or in disease. Metabolic engineering techniques are also finding important uses in tissue engineering, where they can be used to monitor the metabolic response of cells and tissues to perturbations in the environment and rationally design culture media that enhance cell function and proliferation.

In metabolic engineering, the notion of cellular metabolism as a network is of central importance. Also, fundamental to metabolic engineering is the idea that metabolic processes, systemic or cellular, are coupled and as such cannot be considered separately. The major metabolic pathways (e.g., glycolysis, gluconeogenesis,

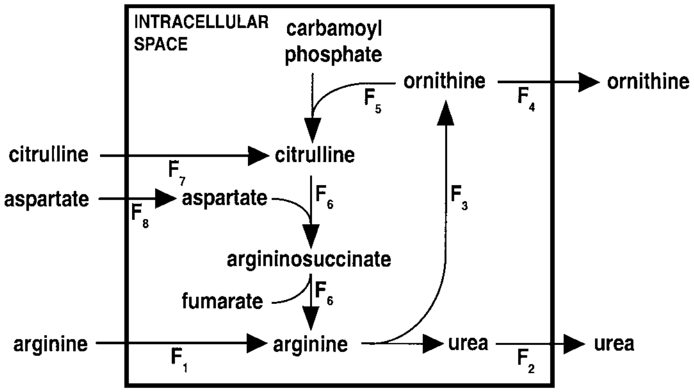
pentose phosphate, urea cycle, tricarboxylic acid cycle, fatty acid synthesis and oxidation) are interrelated through common precursors and metabolic intermediates. Thus, an enhanced perspective of metabolism and cellular function can be obtained by considering a framework that incorporates all the major participating reactions, rather than a few isolated ones. Two methodologies for the characterization and analysis of cell metabolism that are especially useful for the analyses of metabolic abnormalities in human disease are metabolic flux analysis and metabolic control analysis.

Metabolic flux can be defined as the net rate of conversion of one metabolic precursor to a product. Metabolic flux analysis refers to the calculation of fluxes through metabolic pathways. Two techniques are primarily used for flux determination: (1) mass isotopomer analysis, and (2) extracellular metabolite balance models. Mass isotopomer analysis has been used extensively to quantitate fluxes in mammalian cells and tissues including brain, heart, and liver. In this approach, the body is fed, or the isolated tissue is perfused with, substrates labeled with stable isotopes (e.g., ^{13}C). A different labeling pattern of metabolites in the blood, perfusate, and/or tissue extract arises depending on the pathways utilizing these substrates. The labeling patterns are experimentally determined by nuclear magnetic resonance or mass spectroscopy. These labeling patterns are analyzed in conjunction with a mathematical model to calculate the fluxes through the various pathways which best account for the observed labeling patterns. Although isotopomer analysis is a powerful and generally noninvasive method, stably labeled compounds and the instruments required to determine the isotopomer distributions of key metabolites are relatively expensive.

Material balances of whole-body macronutrients have been used since the 19th century for evaluating bulk material processing with the body (e.g., to study the conversion of carbohydrates to fat). Material balances have since evolved into metabolic flux balance models that can predict intracellular fluxes in complex metabolic networks. This methodology utilizes a stoichiometric model that describes the major intracellular reactions at steady state. Extracellular fluxes, which correspond to rates of consumption/production of extracellular metabolites, are experimentally determined, and intracellular fluxes are calculated based on the stoichiometric constraints of the intracellular reaction network. This approach has been used extensively to study and improve strains of microorganisms (bacteria and yeasts) of significance in biotechnology. As of now, applications of metabolic flux balance models to mammalian cell systems have been more limited but are gaining in popularity.

The starting point in this analysis is the construction of a list of steady-state material balance equations to describe the conversion of substrates to metabolic products for the biochemical system of interest. For example, if one considers a simplified scheme of amino acid metabolism in liver, one can write a set of steady-state material balance equations that represent the flow of metabolites through the network (Fig. 9). The equations contain measurable quantities (these are marked with an asterisk) which are the rates of consumption/production of extracellular metabolites. The concentrations of strictly intracellular metabolites (e.g., argininosuccinate) are assumed to be constant. In this particular case, we have eight fluxes to be determined, five of which are measurable (F_1^*, F_2^*, F_4^*, F_7^*, F_8^*). The five equations listed here, which relate these fluxes to each other, can be reduced to four independent equations. Thus, the system can be solved to yield the three unknown intracellular fluxes (F_1, F_2, F_4). Because the system is overdetermined, it provides an internal check for consistency of the data with each other and the assumed biochemistry.

While this method is very useful, there is a limit to the extent to which complex metabolic networks can be



Mass balance equations:

$$F_1^* + F_6 = F_2^*$$
$$F_2^* = F_3$$
$$F_3 = F_4^* + F_5$$
$$F_5 = F_6 + F_7^*$$
$$F_8^* = F_6$$

FIGURE 9 System of mass balance equations describing the flow of metabolites through the urea cycle of hepatocytes. Measurable fluxes are labeled with a star.

elucidated by using measurements of extracellular products alone. For example, flux distribution at split points that converge at another point of the network cannot be resolved. Another limitation is that only net fluxes are determinable, while isotopic methods can sometimes resolve the rates of the forward and the backward reactions. This methodology may be particularly useful when used in combination with stable isotopes to provide fluxes that cannot be directly determined by the isotopomer analysis. Metabolic flux analysis, once validated for the particular case under study, is potentially very useful as it is noninvasive and cost effective.

Another important aspect of the metabolic network that can be investigated by metabolic engineering techniques is the "rate-controlling" enzymes of the pathway (i.e., the enzymes governing flux in the metabolic network). Over the past 30 years, several theoretical frameworks for this type of analysis have been developed. Of these, one of the most widely used is metabolic control analysis. Metabolic control analysis aims at quantifying the control that individual or groups of enzymes exert on the flux through a particular pathway by studying the response of the system to changes in nutrient levels and other factors that alter the activity of specific enzymes in the network. This analysis is generally quite difficult to perform experimentally and is often based on many assumptions; however, it provides valuable insight into the mechanisms governing metabolic adaptation to changes in the environment and a rational basis for genetically engineering cells to perform specific functions.

4. Effects of Mechanical Forces on Cells and Tissues

The environment in which cells are cultured has traditionally been defined by the presence of soluble factors such as hormones and growth factors, and the chemical properties of the surface on which they adhere and grow. In addition, certain physical forces may influence cellular function and may be used as tools to induce specific phenotypes in cells. For example, it is believed that mechanical loading plays an important role in the synthesis and deposition of extracellular matrix by cells in load-bearing tissues such as cartilage and bone *in vivo*. Thus, incorporating mechanical loading schemes in the culture environment such as cyclical compression may be beneficial. Furthermore, the mechanical loading apparatus can be coupled to sensors to provide a continuous assessment of the mechanical parameters (compressive strength and module of elasticity) of the developing tissue. Another example is the effect of fluid shear stress and uniaxial stretch, both of which induce vascular endothelial cell elongation and alignment, as well as the secretion of vasoactive compounds. Cyclic stretch has also been used to promote the fusion of muscle myoblasts (muscle progenitor cells) into contractile myotubes containing parallel myofibers aligned in the direction of the applied force.

C. Transport Phenomena in Tissue Engineering

1. Cell Migration

Cell migration is often a critically important step in many applications of tissue engineering. For example, biomaterial implants used for nerve regeneration, cartilage, and skin wound healing require that the host's cells migrate into the matrix implant. Cell migration speed depends on a complex balance between cell tractional forces and the stickiness of the matrix to the cell. The highest cell speeds are obtained at intermediate attachment strengths, which allow the leading edge of the cell to anchor itself to the surface while the receding edge comes off the surface. A substrate with low adhesiveness does not allow the cell to form any anchors to the surface that can resist cell tractional forces, which results in poor migration. Similarly, cells "glued" to a surface that is too sticky are not able to move forward because cell–substrate bonds at the receding edge of the cell cannot be broken.

The determination of cell-migration parameters is important in order to predict the speed at which cells can invade a tissue construct. The prediction of cell migration behavior based on the knowledge of cell mechanics, interaction of cell receptors with appropriate ligands on the extracellular matrix, and function of the cytoskeleton is possible, but complicated, and involves difficult measurements. On the other hand, phenomenological cell-migration parameters can be determined via analysis of single cell trajectories or cell concentration profiles of cell populations in specific devices which allow for the visualization of cells during the migration process. The most simple model to analyze single cell trajectories is the persistent random-walk model (Fig. 10). This model assumes that cells are not restricted in their range of movement (within the duration of the experiment), they can move in any direction with equal probability, and, once they move in a certain direction, they exhibit a characteristic persistence time before they change direction. The parameters of the model, persistence time (P) and cell speed (S), can be used to calculate an equivalent diffusivity coefficient (also called *random motility coefficient*), which is a measure of the propensity of a cell population to spread:

$$D = \frac{S^2 P}{n} \tag{5}$$

where n is the number of dimensions ($n = 2$ for a surface, $n = 3$ for a gel) where the migration occurs.

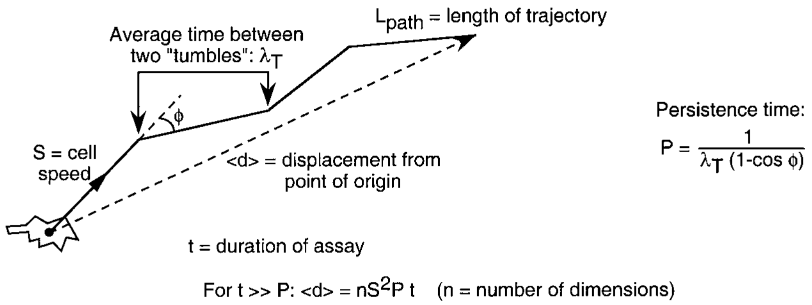


FIGURE 10 Definition of parameters that characterize single cell migration.

It is important to note that these parameters are not dependent on the geometry of the system used to measure them and thus can be used to predict cell migration in other geometries. This model chiefly applies to two-dimensional surfaces; however, it can be extended to three-dimensional matrices, in which case the effective pore size of the matrix, which may create a hindrance to the migration process, needs to be taken into account. Cell migration in a specific direction can be promoted by micropatterning tracks on a surface, which prevents cells from wandering away from the desired direction, and in three dimensions by using materials exhibiting oriented pores and/or fibers.

The above discussion relates to the process of random migration where cells do not move in a preferential direction. It is often the case, however, that soluble and insoluble factors causing cells to move in a preferential direction are present. Soluble agents that "attract" cells are called *chemotactic*, while those immobilized in the extracellular matrix are called *haptotactic*. In chemotactic or haptotactic migration, a third parameter must be determined to capture the directional preference of the migration process. This parameter is the chemotactic index (*CI*), which can be determined experimentally from single cell trajectories by the equation:

$$CI = \frac{\langle d \rangle}{L_{\text{path}}} \tag{6}$$

where $\langle d \rangle$ is the distance of the cell from the point of origin at the beginning of the experiment, and L_{path} is the length of the path used by the cell to achieve the displacement $\langle d \rangle$. The population-relevant parameter that describes chemotaxis is the chemotaxis coefficient χ, which is calculated by the following expression:

$$\chi = \frac{S \cdot CI}{\nabla L} - \frac{1}{n}\left[\frac{d\ln P}{dL} - \frac{d\ln S}{dL}\right] \tag{7}$$

where L and ∇L are the concentration and spatial gradient, respectively, of chemoattractant or haptotactic factor. Because chemotactic and haptotactic factors may increase cell speed, and thus increase migration via a "chemokinetic" or "haptokinetic" effect, the last term in Eq. (7) includes a correction for this effect.

The values of D and χ are the constitutive parameters describing cell migration in a variety of both *in vitro* and *in vivo* systems. The expression analogous to Fick's first law of diffusion for cell flux is

$$J = -D\frac{\partial C}{\partial x} + C\left(-\frac{dD}{2dL} + \chi\right)\frac{\partial L}{\partial x} \tag{8}$$

and a cell concentration profile in any system can be derived via the continuity equation:

$$\frac{\partial C}{\partial t} = -\frac{\partial J}{\partial x} \tag{9}$$

In real cases, Eq. (9) may need to be solved in conjunction with appropariate transport equations for the chemoattractant or haptotactic factor, which may be time varying.

As briefly discussed earlier, cells migrating on substrates exert forces that allow them to move. As a result, cells on surfaces or inside gels that are compliant can significantly alter the shape of the material. Quantitative analyses and mathematical descriptions of these phenomena allow prediction of how the cell-material construct changes shape over time. A well-known example of cell-mediated contraction is the fibroblast-populated collagen lattice, which forms the basis for some of the currently used tissue engineered skin grafts. The fibroblast-populated collagen lattice is generated by mixing fibroblasts with a chilled solution of collagen in physiological buffer, followed by exposure to 37°C to induce the gellation of the collagen. If the gel is not anchored to any surface, fibroblasts embedded in a collagen gel cause the contraction of the gel in an isotropic fashion. The contraction process can be controlled to a certain extent by mechanically restricting the motion along certain directions, which also induces a preferential alignment of the collagen fibers as well as the cells within it, which results in a nonisotropic connective tissue equivalent. Preferential alignment of cells may be important in specific applications, such as in tissue

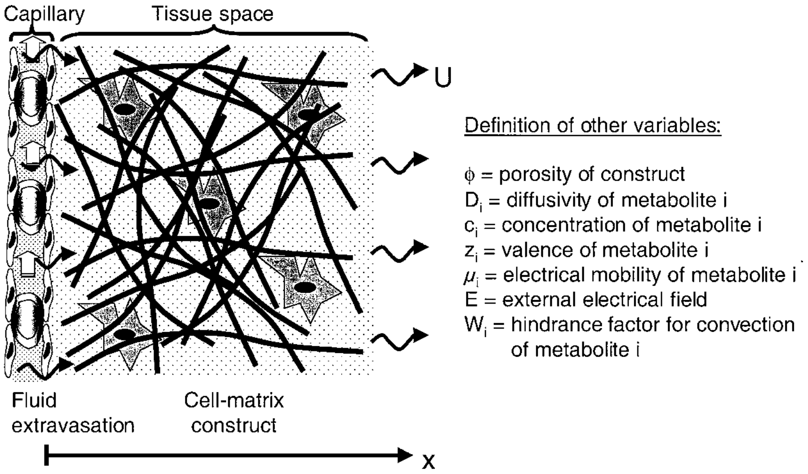


FIGURE 11 Transport of metabolites into a tissue construct implanted next to a blood vessel. *U* is the velocity of fluid extravasated into the tissue; see text for additional explanations.

engineered blood vessels. A potential design of such vascular grafts would have endothelial cells on the inside surface of the graft aligned with the direction of blood flow, as is observed *in vivo* in arteries. On the other hand, the appropriate direction of smooth muscle cells within the blood vessel wall would be along the circumference of the vessel in order to be able to perform their natural function of modulating the diameter of the vessel.

2. Metabolite Transport

In normal tissues, the circulatory system brings in nutrients and removes waste products. Typically, no cell *in vivo* is farther than about 100 μm, or even sometimes less, from a blood vessel. Thus, transport by diffusion does not have to occur over distances beyond 100 μm. Engineered tissue constructs are mostly devoid of any vascular system, and although there may be culturing methods allowing transport by convection throughout the cell mass, as described later in the bioreactor section, after implantation the engineered tissue is no longer perfused and the situation remains so until vascularization by angiogenesis occurs from the surrounding host's tissues. Vascularization *in situ* can be accelerated by implanting tissues that release angiogenic factors, such as fibroblast-derived growth factor and vascular endothelial growth factor. Although tissue constructs may include vascular endothelial cells, it is not yet possible to create three-dimensional vascular networks *in vitro* that are patent.

Transport through tissues can be modeled using the basic transport equations used in nonliving systems and can be useful to predict the concentration profiles of metabolites throughout engineered tissues. In the following presentation, we apply these equations with the goal of providing design criteria for tissue constructs. We consider a tissue construct of thickness X implanted *in vivo*, assuming that this construct is avascular but surrounded with blood vessels from the hosts's tissue (Fig. 11). The fundamental equation describing the flux N of a particular species i is given by Fick's law of diffusion, to which terms to account for the electrical migration of species and convection are added:

$$N_i = \phi\left\{-D_i\frac{\partial c_i}{\partial x} + \frac{z_i}{|z_i|}\mu_i c_i E\right\} + W_i c_i U \qquad (10)$$

To facilitate the comparison between systems of different geometries and scales, we next rewrite the previous equation using dimensionless quantities. The Péclet number is a dimensionless number defined as the ratio of nondiffusive transport (convection and electrical migration) to transport by diffusion in a particular system:

$$Pe = X\left(\frac{W_i U + \phi\frac{z_i}{|z_i|}\mu_i E}{\phi D_i}\right) \qquad (11)$$

Substituting into the flux equation yields:

$$N_i = \phi D_i\left(-\frac{\partial c_i}{\partial x} + \frac{Pe}{X}c_i\right) \qquad (12)$$

We can further simplify this equation by defining the following dimensionless variables:

$$N_i^* = \frac{N_i}{\phi D_i}, \qquad c_i^* = \frac{c_i}{X}, \qquad x^* = \frac{x}{X} \qquad (13)$$

which yields a new and simplified flux equation containing only the Péclet number as the unknown or "adjustable" parameter, the value of which depends on the system under consideration:

$$N_i^* = \left(-\frac{\partial c_i^*}{\partial x^*} + Pe\, c_i^*\right) \qquad (14)$$

The use of the dimensionless quantities facilitates comparison between systems with different geometries and scales. For example, although two systems may be different in several respects, they will behave similarly as far as metabolite transport if they have similar Péclet numbers. Furthermore, the Péclet number gives an instant idea as to which major transport mechanisms operate in the system under study. If $Pe < 1$, diffusional transport dominates, while if $Pe > 1$, convective transport and electrical migration are more important.

We now consider the typical case of a tissue engineered construct implanted *in vivo*. Although the precise values of the transport parameters are not necessarily known, it is often a useful exercise to perform an order of magnitude analysis to determine the approximate contribution of each term in the transport equation. For this purpose, we start with Eq. (12) and propose that a reasonable estimate for the concentration gradient $\partial c_i/\partial x$ is c_i/X. The flux equation becomes:

$$N_i = \phi D_i\left(-\frac{\partial c_i}{\partial x} + \frac{Pe}{X}c_i\right) \approx \phi D_i\left(\frac{c_i}{X} + \frac{Pe}{X}c_i\right) = \frac{\phi D_i c_i}{X}(1 + Pe) \tag{15}$$

Most nutrient transport to the implant initially comes from the surrounding capillaries. These capillaries continuously leak plasma into the tissue space because the pressure inside capillaries is greater than in the tissue. Typical measured values for the capillary filtration coefficient and pressure gradient are 0.035 cm^3/min/mm Hg/100 g tissue and 27 mm Hg, respectively. Assuming capillaries are distributed evenly 100 μm (or 0.01 cm) apart, on average each capillary occupies (0.01 cm)^3 = 10^{-6} cm^3. Since 100 g tissue occupy a volume of about 100 cm^3, we can estimate the flow rate and velocity of fluid exiting capillaries and flowing through the implant:

$$Q = 0.035\frac{\text{cm}^3}{\text{min}\cdot\text{mmHg}\cdot 100\,\text{cm}^3\,\text{tissue}} \times 27\,\text{mmHg} \times \frac{1}{60\text{s}} \times \frac{10^{-6}\text{cm}^3}{\text{capillary}} = 1.2\times 10^{-8}\frac{\text{cm}^3}{\text{capillary}\cdot\text{s}}$$

The average surface area perfused by each capillary is 0.01 cm×0.01 cm, thus the velocity is

$$U = \frac{1.2\times 10^{-8}\text{cm}^3/\text{s}}{(100\times 10^{-4}\text{cm})^2} = 1.6\times 10^{-4}\text{cm/s}$$

We next assume that the implant is 0.1 cm thick and highly porous (as would be the case for a collagen gel, for example), so that $\phi = 1$, with a mean pore size of >10 μm, which is much larger than the molecular size of transported molecules, so that $W_i = 1$. Using these values, we calculate the Péclet number and corresponding flux rate for oxygen, glucose, and a peptide hormone, insulin, in the tissue construct (Table VI). For the small metabolites oxygen and glucose, Péclet numbers are close to 1, which indicates that diffusive and convestive transports have equal contributions. Insulin, by virtue of its molecular size, has a lower diffusivity, thus its transport is more dependent on convection. One can notice that although oxygen diffusivity is at least one order of magnitude greater that that of glucose, the transport of oxygen is almost two orders of magnitude slower. This is due to the fact that oxygen has a very low solubility in water and physiological fluids. Thus, although oxygen diffuses rapidly, it is not possible to create large gradients to provide the necessary driving force for its transport; for this reason, oxygen transport is almost always the main factor limiting the size and cell density of tissue engineered constructs.

TABLE VI Transport Parameters for a Few Metabolites Important for the Function of Tissue Engineered Constructs Implanted *in vivo*

Metabolite	D_i (cm^2/s)[a]	C_i (mM)[b]	Pe	N_i (μmol/cm^2/s)
Oxygen	2×10^{-5}	0.1	0.8	4×10^{-5}
Glucose	9×10^{-6}	5	1.8	120×10^{-5}
Insulin	1.5×10^{-6}	3×10^{-8}	11	5×10^{-12}

[a] D_i are typical values measured at 37°C (the body temperature).
[b] C_i are typical values found in biological fluids *in vivo* (for oxygen, arterial blood levels were used).

To obtain an estimate of the maximum thickness of a tissue engineered implant based on transport considerations, we must balance metabolite delivery with consumption by the cells in the construct. For this purpose, we use the mass balance or continuity equation, which is generally written as:

$$\frac{\partial c_i}{\partial t} = -\frac{\partial N_i}{\partial x} + \sum_i (G_i - R_i) \tag{16}$$

where G_i and R_i represent the generation and consumption rates of metabolite i in the construct, respectively. Substituting the expression for the flux N_i yields:

$$\frac{\partial c_i}{\partial t} = D_i\left\{\frac{\partial^2 c_i}{\partial x^2} + \frac{Pe}{X}\left(\frac{\partial c_i}{\partial x}\right)\right\} + \sum_i (G_i - R_i) \tag{17}$$

Let us now consider the specific case of oxygen transport. Furthermore, to simplify the calculations, we assume that convective transport is negligible so that $Pe = 0$. This would be a "worst-case scenario" where normal tissue perfusion is disrupted due to the surgical trauma caused by the implantation procedure itself, or the implant is encapsulated into a membrane (i.e., to protect implanted cells from the host's immune system) which does not allow

convective transport into the implant. Furthermore, there is no generation of oxygen by the tissue, so $G_i = 0$. The consumption of metabolites by cells often follows first-order kinetics at low concentrations (i.e., the consumption rate is proportional to the concentration of metabolite) and progressively becomes zero order as the concentration increases. At one point, the cells are "saturated" and cannot take up more. This behavior is often described by Michaelis–Menten kinetics:

$$R_i = \frac{V_{\max} c_i}{K_M + c_i} \tag{18}$$

where $V_{\max}$ is the maximal uptake rate and K_M is the metabolite concentration at which the uptake rate is half-maximal. Substituting this expression into the continuity equation and taking into account other assumptions described above yield:

$$\frac{\partial c_i}{\partial t} = D_i \frac{\partial^2 c_i}{\partial x^2} - \frac{V_{\max} c_i}{K_M + c_i} \tag{19}$$

If we next assume steady state, we obtain:

$$0 = D_i \frac{\partial^2 c_i}{\partial x^2} - \frac{V_{\max} c_i}{K_M + c_i} \tag{20}$$

One can integrate this expression with the following boundary conditions:

$$x = 0, \qquad c_i = C_0 \tag{21}$$

$$x = X, \qquad \partial c_i / \partial x = 0 \tag{22}$$

to yield the concentration profile throughout the system. However, rather than looking at the whole concentration profile, we really only want to know if at any point within the system there will be a significant depletion of metabolite. As a rule of thumb to estimate when a cell is starving, the condition $c_i = K_M$ is often used. Figure 12 shows the results for the maximum thickness X of a construct without c_i going below K_M anywhere in the construct. Using this chart to estimate the thickness of an implantable construct containing liver parenchymal cells (hepatocytes), we find that a construct containing even a relatively low density of cells (10^7 cells/cm^3) cannot have a thickness exceeding about 500 μm (Table VII). At tissue-level cell densities of 10^8 cells/cm^3, that thickness can be as low as 100 μm, which is consistent with the *in vivo* density of capillary vessels.

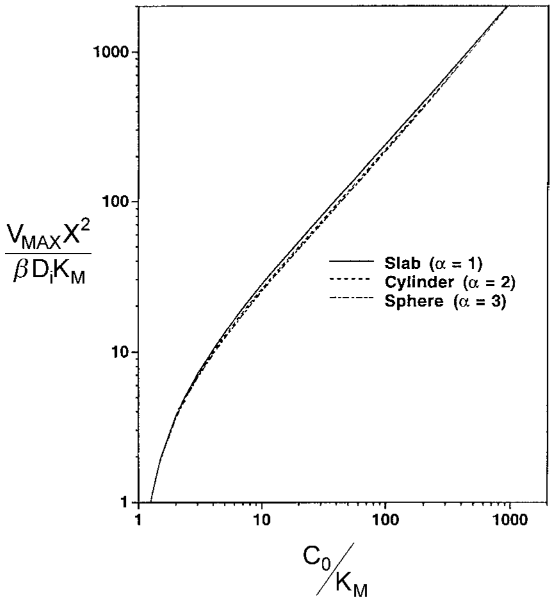


FIGURE 12 Correlation for the maximum thickness of a tissue construct to avoid nutrient depletion.

3. Bioreactor Technologies

For certain applications it is necessary to maintain a large number of cells that transform an input of reactants into an output of products. This is the case for the bioartificial liver or pancreas and more recently for the production of blood cells from hematopoietic tissue. These systems require maintenance of the function of a large number of cells in a small volume. For example, a hypothetical bioartificial liver device possessing 10% of the detoxification and protein synthesis capacity of the normal human liver (a rough estimate of the minimum processing and secretory capacities that can meet a human body's demands) would contain a total of 10^{10} adult hepatocytes. Thus, to keep the total bioreactor volume within reasonable limits (1 L or less), 10^7 cells/mL or more are required. For comparison, the normal human liver contains approximately 10^8 hepatocytes/mL. Three main types of bioreactor design have been considered in tissue engineering: (1) suspension culture methods using microcarriers, (2) cells immobilized in hollow-fiber systems, and (3) suspension culture in rotating wall vessel bioreactors.

a. Microcarrier-based systems. Microcarriers are one of the first methods used for supporting large-scale mammalian cell culture. Microcarriers are small beads

TABLE VII Maximum Thickness (*X*) of Tissue Constructs Estimated by Order of Magnitude Transport Analysis[a]

Geometry	10^7 cells/cm^3 X (μm)	10^8 cells/cm^3 X (μm)
Slab	300	95
Cylinder	430	135
Sphere	520	165

[a] These estimates are based on the assumptions that that surface of construct is exposed to arterial oxygen levels ($C_0 = 100$ nmol/cm^3), hypoxic damage occurs when $c_i = K_M$, and for hepatocytes $V_{MAX} = 0.4$ nmol/10^6 cells/s, $K_M = 0.6$ nmol/cm^3.

(usually less than 500 μm in diameter) with surfaces treated to support cell attachment. These beads are then maintained in suspension in medium using very low stirring speeds in order to avoid mechanical cell damage, either due to shearing forces in the liquid or due to bead–bead collisions. The surface area available per microcarrier can be increased by using porous microcarriers, where cells can migrate and proliferate within the porous matrix as well as on the microcarrier surface; furthermore, cells within the microcarrier are protected from mechanical damage. To attach cells to microcarriers, isolated cells are mixed with microcarriers in suspension. The protocol requires careful optimization of the number of cells per bead, mixing velocity (intermittent mixing may be necessary until cells are firmly attached), and supply of oxygen, which is necessary for cell attachment as the cells require energy in order to spread onto a substrate.

Bioreactor configurations using microcarriers include packed and fluidized beds. A packed bed of microcarriers consists of a column filled with microcarriers with porous plates at the inlet and outlet of the column to allow perfusion while preventing microcarrier entrainment by the flow. Reactor volume is proportional to the microcarrier diameter, thus it is advantageous to reduce the microcarrier size as much as possible. However, packed beds with small beads may clog and the cells may have a tendency to accumulate in the channels between the microcarrier surfaces. Total flow rate is mainly dependent on cell number and the nutrient uptake rate of the cells. Because oxygen is usually the limiting nutrient, the medium flow rate through the reactor is found using the following equation:

$$\text{Flow rate} = \frac{\text{O}_2\ \text{Consumption per cell} \times \text{Cell number}}{\text{O}_2\ \text{Concentration in medium}} \tag{23}$$

The aspect ratio of the bed (height/diameter) determines the fluid velocity through the packed bed according to the equation:

$$\text{Fluid velocity} = \frac{\text{Flow rate}}{\text{Cross-sectional area}} \tag{24}$$

and is adjusted so that the magnitude of fluid mechanical forces (proportional to the aspect ratio) within the bed is below damaging levels (Fig. 13). Fluidized beds differ from packed beds in that the perfusing fluid motion maintains the microcarriers in suspension. Packed-bed systems have been shown to support cell densities exceeding 10^8 cells/mL when using microporous microcarriers (500 to 850 μm in diameter). In addition, packed beads (1.5-mm diameter) have been used to entrap aggregates of hepatocytes. The latter application was shown to maintain a relatively stable level of albumin secretion (a liver-specific product) for up to 3 weeks.

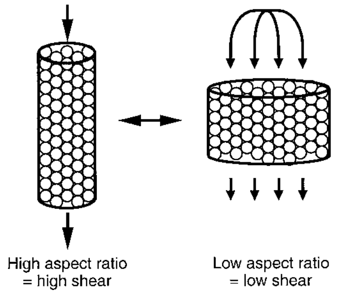


FIGURE 13 Two possible configurations for a packed bed bioreactor of equal volumes.

b. Hollow-fiber systems. The hollow-fiber system is the most widely used type of bioreactor used in tissue engineering and in artificial organ development. It consists of a shell traversed by a large number of small-diameter tubes (Fig. 14). The cells may be placed within the fibers in the intracapillary space or on the shell side in the extracapillary space. The compartment that does not contain the cells is generally perfused with culture medium or the patient's plasma or blood. The fiber walls may provide the attaching surface for the cells and/or act as a barrier against the immune system of the host. Microcarriers have also been used as a way to provide an attachment surface for anchorage-dependent cells introduced in the shell side of hollow-fiber devices. Hollow-fiber systems can be designed to be implanted as vascular shunts, but may also be perfused with the patient's blood or plasma extracorporeally.

There are many studies on how to determine fiber dimensions, spacing, and reactor length; however, commercially available units come in a relatively limited number of sizes, usually with inner fiber diameters of 500 μm or more. Several reports in the literature describe the use of hollow-fiber systems in the development of a bioartificial pancreas, which place the islets on the shell side, while perfusing the fibers with the animal's plasma or blood. The fibers can be made relatively non-thrombogenic and of porosity sufficiently small as to avoid immune attack of the cells inside the shell. One difficulty with this configuration is that interfiber distances in the hollow-fiber device are not well controlled, so that regions within the shell space receive too little nutrients.

It may be advantageous to place cells in the lumen of small fibers because the diffusional distance between the shell (where the nutrient supply would be) and the cells is essentially equal to the fiber diameter, which is easier to control than the interfiber distance. In one configuration, cells have been suspended in a collagen solution and injected into the lumen of fibers where the collagen is allowed to gel. Contraction of the collagen lattice by the cells even creates a void in the intraluminal space, which can be perfused with hormonal supplements, etc. to enhance

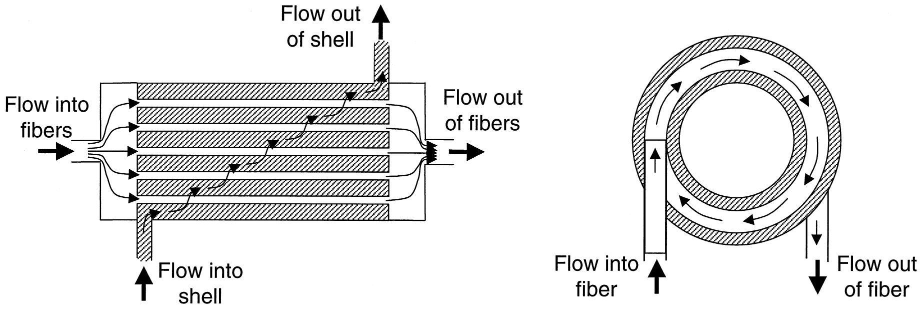


FIGURE 14 Two common types of hollow-fiber bioreactors. In the reactor shown on the left, cells may be placed either in the shell (gray color) or intrafiber space. In the reactor shown on the right, cells are typically placed in the shell space.

the viability and function of the cells, while the patient's plasma would flow on the shell side. Such a configuration has been described for the construction of a bioartificial liver using adult hepatocytes.

c. Rotating vessel wall bioreactor. The rotating wall vessel bioreactor, a relatively new type of bioreactor system used in biotechnology, was originally developed and patented by the U.S. National Aeronautic Space Agency (NASA) to study the behavior of cells and tissues under conditions simulating low gravity on Earth. This bioreactor consists of a chamber entirely filled with culture medium containing the cells, tissue constructs, or even actual tissue explants in suspension. The chamber is rotated on a horizontal axis at a speed that approximately matches the terminal settling velocity of the cells or tissues in suspension such that they establish a fluid orbit (Fig. 15). The cells or tissues therefore never hit the bottom of the reactor or touch any of its inner surfaces. An important feature of this system is that oxygen is delivered via gas-permeable silicone membranes; no sparging of gas is necessary. The design ensures uniform hydrodynamic conditions within the bioreactor without the use of impellers, which have been shown to cause significant cell damage in other types of bioreactors. Interestingly, data gathered on recent space missions do suggest that cells and tissues cultured in this bioreactor develop and grow as they would in low-gravity environments.

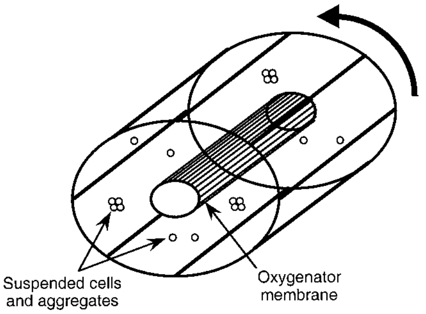


FIGURE 15 Design of a rotating wall vessel bioreactor. Cells are suspended in medium, which fills the vessel until no air bubble is left. Oxygen is delivered via a silicone membrane in the center of the vessel. The vessel rotates at a relatively low speed (~30 rotations/min) to prevent settling of cells.

Over 50 types of cells, tissue constructs, and even tissue explants have been cultured in these bioreactors, which appear to be ideally suited to promote the expression of tissue-specific functions in the cultured cells and preserve the three-dimensional morphological characteristics of the native tissue. Thus, this system should be useful to create and maintain bioartificial tissues to be subsequently implanted *in vivo*. On the other hand, these devices would not be appropriate for tissue engineering applications requiring a combination of very high cell densities and very low liquid hold-up volumes, such as in the case of extracorporeal bioartificial livers.

D. Morphogenesis of Engineered Tissues

The quantitative difference between cell–substrate and cell–cell adhesion strength on a rigid surface dramatically affects the organization of cells on the substrate. A thermodynamic view of the problem suggests that the overall system (consisting of the cells and the extracellular support) ultimately reaches an equilibrium state when the surface free energy is minimized. According to this concept, the existence of large cell–substrate adhesion forces relative to cell–cell adhesion forces prevents cell–cell overlapping (Fig. 16). In contrast, the opposite situation would lead to cell clumping or multilayered growth on the substrate. This prediction is in agreement with the observation of cellular aggregate formation when hepatocytes are plated on a nonadherent surface as opposed to a highly adherent surface such as type I collagen.

Heterotypic cell systems or "co-culture" systems have been used for the production of skin grafts, in long-term cultures of hepatocytes, and in long-term cultures of mixed bone marrow cells. These systems take advantage of the

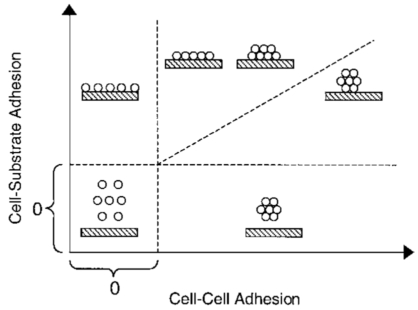


FIGURE 16 Possible configurations of cells on a flat substrate. In the absence of any adhesion, cells remain in single cell suspension (bottom left quadrant). Increasing substrate-cell adhesion causes cells to stick to the surface. Increasing cell–cell adhesion from that point causes the cells to attach to each other and form monolayers. Increasing cell–cell adhesion further promotes aggregation of the cells on the substrate.

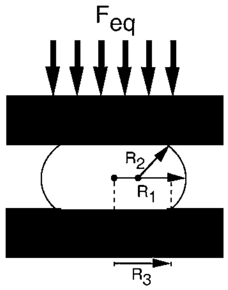


FIGURE 17 Deformation of cell aggregate during compression.

trophic factors (for the most part unknown) secreted by "feeder" cells. Greater use of different cell types used in co-culture will enable engineered cell systems to closely mimic *in vivo* organization, with potential benefits including increased cell function and viability and greater range of functions expressed by the bioartificial tissue. The organization of multicellular three-dimensional structures may not be obvious. Provided that the adherence of homotypic and heterotypic interactions is known, a thermodynamic analysis similar to that used to describe the morphology of a pure cell culture on a surface can be used to predict how cells will organize in these systems. The process of cell–cell sorting in multicellular systems may be altered by changing the composition of the medium or altering the expression of proteins mediating cell–cell adhesion via genetic engineering.

Predicting the organization of simple multicellular systems containing more than one cell type is possible if the relative cohesiveness of each cell type with respect to each other is known. The cohesiveness of a tissue is exactly the same as a tissue surface tension analogous to a liquid surface tension. The tissue surface tension σ has been measured in homogenous three-dimensional cellular aggregates by measuring the compression force required to deform an aggregate into a flattened droplet shape (Fig. 17).

The force exerted by the cell aggregate on the compression plate is given by the Laplace equation:

$$\frac{F_{eq}}{\pi R_3^2} = \sigma\left(\frac{1}{R_1} + \frac{1}{R_2}\right) \tag{25}$$

where F_{eq} is the force measured after sufficient time has been allowed for the aggregate, which behaves as a viscoelastic liquid, to relax.

When tissues A and B are combined, they will reorganize depending on the relative values of σ. At equilibrium, they will reach a predictable configuration that minimizes the surface free energy. In the case where each cell has uniform stickiness on its entire surface, the possible configurations are illustrated in Fig. 18.

This theory has been tested and verified extensively using aggregates of embryonic cells. Furthermore, by changing the level of expression of surface adhesion molecules via hormonal induction or genetic engineering, it was possible to alter the organization of such cell aggregates in a predictable fashion. It is important to note, however, that although the theory was derived for a closed system at equilibrium, all of the cultured cell systems examined were in fact open systems because the cells were dissipating energy provided by the culture medium. Thus, care should be taken when using this approach to analyze more complex cases. In addition, further refinement to the application of the equations will be required in cases where cell adhesiveness is not uniformly distributed. This situation is common to most epithelial cells, which exhibit poor to negligible cell–cell adhesiveness on the apical surface, and as a result tend to form tubular cell structures.

E. Summary

Tissue engineering is the construction of bioartificial tissues *in vitro* as well as the *in vivo* alteration of cell growth and function via implantation of suitable cells isolated from donor tissue and biocompatible scaffold materials. Biomaterials for tissue engineering must have controlled surface chemistry, porosity, and biodegradability in order to promote optimal cell adhesion, migration, and deposition of endogenous extracellular matrix materials by the cells. Strategies to switch cells between growth and differentiation, which tend to be mutually exclusive, are used in order to provide a large cell mass that can perform specific differentiated functions required for the tissue construct. Combinations of cells and materials have the ability to reorganize themselves based on the strength of adhesion between cells and substrate and among the various cell types present in the tissue construct. Finally, tissue constructs must be intimately integrated into the

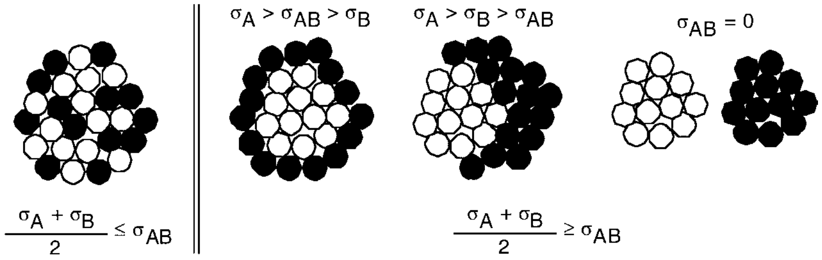


FIGURE 18 Possible equilibrium configurations of two cell types A and B mixed together depending on the relative tissue surface tension of each tissue. Cells will not remain mixed unless the adhesion force of A to B exceeds the average cohesion of tissues A and B (left-most case). Otherwise, cells segregate and the most cohesive tissue will tend to remain in the center while the other tissue spreads around it.

host's vascular system in order to provide efficient nutrient supply and waste removal.

III. APPLICATIONS OF TISSUE ENGINEERING

A. Connective Tissues

1. *In Vitro* Construction of Connective Tissues

Connective tissues can be reconstructed *in vitro* by incorporation of connective tissue cells within a porous biomaterial or a loose network of extracellular matrix components. The resulting geometry and organization of the tissue equivalent are similar to those of the parent tissue *in vivo*. Several different types of connective tissues have been built using this approach (Table VIII). The mechanical properties of the original biomaterial are dramatically altered by the embedded cells, due either to cell-generated forces causing contraction of the matrix material or deposition of extracellular matrix generated by the cells themselves. This remodeling is often important for the eventual function of the tissue construct.

The only engineered connective tissue equivalent currently used clinically is bioartificial skin. Bioengineered skin has several applications, including the treatment of burn wounds and nonhealing diabetic and venous leg ulcers, as well as pressure sores. Bioartificial skin consists of a dermal equivalent made by seeding collagen gels or meshes made of biodegradable polymers with dermal fibroblasts. The dermal equivalent may have an overlay of silicone sheeting to prevent evaporative loss of water (a function normally performed by the missing epidermis). After take, the grafted material is suitable for grafting an epidermis, either in the form of a split-thickness skin graft or cultured layer of epidermal cells (keratinocytes). Bioartificial skin is also available as a complete dermal–epidermal composite comprising the dermal equivalent on top of which keratinocytes have been cultured to confluence. Differentiation of the epidermal layer into a functional epidermis with a cornified layer that exhibits a high resistance to chemical damage is induced by exposure to the air–liquid interface during the culturing process. The composites are then grafted onto the wound site in a single operation.

TABLE VIII Examples of Connective Tissues Made *in vitro*

Tissue	Biomaterial	Cell type
Dermis	Collagen	Fibroblast
Vascular media	Collagen, endogenously produced matrix	Smooth muscle cell and fibroblast
Cartilage	Collagen–glycosaminoglycan complex, polylactic–glycolic copolymer	Chondrocyte
Meniscus	Collagen–glycosaminoglycan complex	Fibrochondrocyte
Bone	Calcium phosphate, polylactic–glycolic copolymer	Mesenchymal stem cell from periosteum
Tendon	Collagen	Tenocyte (tendon fibroblast)

Metabolic engineering of connective tissue cells such as fibroblasts and smooth muscle cells via addition of ascorbic acid to the culture medium has been shown to promote the production of large quantities of extracellular matrix, such that the cells produce their own scaffolding material. This is an efficient method to generate sheets of cells that can then be layered on top of each other or rolled around a mandrill to form thicker tissue constructs. Although still in early experimental stages, this approach has been used to form tubes that can withstand significant mechanical stress and is currently being evaluated for the generation of media for bioengineered blood vessels.

Connective tissues that must bear significant loads must exhibit specific mechanical properties. This is the case of bioartificial cartilage, which would be best implanted once its mechanical properties are similar to that of the authentic tissue. Chondrocytes seeded at high density (10^7 cells/cm^3) in agarose gels retain their phenotype and remain viable for up to 6 months. In this system, the

deposition of type II collagen and highly charged proteoglycans, which are primarily responsible for the mechanical properties of native cartilage, is enhanced by subjecting the tissue to cyclical mechanical compression. More recently, chondrocytes seeded in polylactic–glycolic scaffolds and then implanted in ectopic sites *in vivo* were found to generate hyaline cartilage tissue with an overall shape similar to that of the original synthetic matrix. Thus, it may be possible to first generate the tissue of desired properties at ectopic sites and, when ready, implant it at the site requiring intervention.

2. *In Vivo* Regeneration Using Guidance Templates

It is sometimes more convenient to promote tissue regeneration *in situ*, in which case the task of the tissue engineer is to favor wound healing and help the body overcome some of its own limitations with respect to tissue regeneration. The first regeneration templates that became widely available are biodegradable meshes for the treatment of burn wounds. These templates are made of cross-linked collagen–glycosaminoglycan complexes and are applied to the wound site to favor regeneration of the skin dermis. The regenerated surface then provides a suitable substrate for the attachment of skin epidermal cells (keratinocytes), which can be from an autologous skin graft obtained from a donor site elsewhere on the patient or from cultured skin. In animals models, it appears that one of the main benefits of such templates is to slow down wound contraction and favor the production of new tissue resembling skin. The beneficial effect of the template depends on pore size and degradation rate. In humans, the templates appear to favor the production of normal dermis as opposed to disfiguring scar tissue.

Another area of great promise for regeneration templates is to promote the reconnection of severed nerves. The natural regeneration ability of peripheral nerves is limited to about 1 cm. This limitation appears to be chiefly related to the formation of scar tissue, which impedes the axonal regeneration process. To reconnect nerves over longer distances, tubes containing suitable biomaterials that promote growth of axons and inhibit scar-tissue formation are sutured to the ends of the nerve stumps. The material consists of a collagen–glycosaminoglycan composite similar to that used for skin regeneration, except for a faster degradation rate and a smaller pore size (5 μm). In addition, animal studies indicate that regeneration is better when pores are oriented along the longitudinal axis of the tube. Functional results obtained with such nerve regeneration templates in animal models approach those obtained for nerve autografts, the conventional treatment for nerve reconnection.

B. Epithelia and Endothelia

1. Secretory and Transport Functions of Epithelial and Endothelial Cells

Epithelial and endothelial cells separate different compartments in the body; for example, endothelial cells separate the intravascular from tissue space, and intestinal epithelium separates the gut lumen from the inside of the body. They control transport across these compartments, forming a selective barrier that prevents the translocation of certain metabolites while favoring the transport of others, sometimes through energy-dependent processes (especially when the direction of transport is against the concentration gradient). In some cases, epithelial and endothelial cells also perform important secretory functions, such as the release of antithrombogenic factors by endothelial cells and an array of secretory and biochemical functions by liver hepatocytes. Although hepatocytes *in vivo* also perform transport functions and form a separate bile canalicular network, all current approaches to bioartificial liver development essentially ignore this property due to the complexity of reproducing the *in vivo* arrangement of hepatocytes in liver. Furthermore, it has been hypothesized that the most important hepatic functions required for survival involve secretory and biochemical functions that do not require a functional bile canalicular network. A partial list of tissue engineered endothelial and epithelial tissues is given in Table IX.

2. Tissue Constructs Using Epithelial Cells

When the barrier function of the epithelium or endothelium is an important component of the design of the tissue, cells can be cultured on a smooth surface, which allows cells to form a monolayer. The cells then often form tight junctions between themselves which are similar to that found *in vivo*. The barrier function can be assessed via measurement of the electrical conductivity across the monolayer or rate of leakage of proteins as well as other

TABLE IX Examples of Epithelia and Endothelia Made *in vitro*

Tissue	Cell type	Major function (s)
Vascular endothelium	Endothelial cell	Transport and secretion
Cornea	Corneal epithelial cell	Transport
Intestine	Enterocyte	Transport
Liver	Hepatocyte	Secretion
Bladder	Uroepithelial cell	Transport
Skin	Keratinocyte	Transport
Kidney	Kidney epithelial progenitor cell	Transport

relevant compounds. The surface onto which cells are grown is often shaped according to the final application. For example, bioartificial vascular grafts have been produced by seeding endothelial cells onto the luminal surface of small-diameter (6 mm or less) synthetic vascular grafts.

In cases where the ability to control transport is not an important aspect of the function of the bioengineered tissues, a simple way to maintain certain epithelial cells is via seeding in a three-dimensional matrix in a manner similar to that used for connective tissue construction. For example, hepatocytes can be maintained in porous or mesh-type matrices made of a variety of materials including polylactic–glycolic copolymer, alginate, etc., wherein they have a tendency to aggregate. Such aggregates (sometimes called *organoids*) are known to contain cells that have maintained their phenotypic stability. However, for this process to be beneficial, the aggregate size must be controlled to prevent the formation of large aggregates with anoxic cores.

3. Epithelial and Connective Tissue Composites

The long-term survival of endothelial and epithelial cells seeded onto artificial polymers is often problematic due to inflammatory responses and the poor retention of the seeded cells under *in vivo* conditions. Relevant to the case of vascular grafts, it is noteworthy that cultured endothelial cells rapidly become senescent and die *in vitro* when growth factors are removed from the culture medium. Thus, long-term retention of the cultured endothelium often requires the continual presence of such factors in the implant. A solution to this problem, which also provides a more intrinsically biocompatible approach to vascular graft production, is to seed cells on an *in vivo*-like stromal layer releasing the necessary trophic factors. Similarly, keratinocytes require either growth factors or a mesenchymal "feeder layer" of cells in order to survive and grow.

Some epithelial tissues have by the nature of the functions performed an extremely complex topology and organization, which makes it difficult to reproduce faithfully *in vitro*. For example, the intestine has on one side a layer of epithelial cells which selectively transport metabolites into the interstitial tissue space, where a large number of microscopic lymphatic and blood vessels absorb the transported metabolites. The epithelial layer is organized in the form of villi or invaginations to increase the surface area of exchange. Although it may be at some point possible to reproduce these structures "from scratch," it turns out that some of these features can be generated by the cells themselves if placed in the correct environment, in which case they display an amazing ability to reorganize themselves into functional tissues. Recent efforts in the development of bioengineered neointestine use cellular aggregates isolated from intestinal tissue which are seeded onto polymer meshes shaped in a tubular form and anastomosed to the small bowel in animal models. The aggregates develop and form villus structures lined with a columnar epithelium reminiscent of the *in vivo* counterpart, and in some areas a subjacent connective tissue containing smooth muscle cells develops. Anastomosis to the existing bowel and bowel resection in the host receiving the implants have significantly improved morphogenesis and differentiation of the implant.

C. Endocrine Tissues

Major efforts in this area focus on the development of a bioartificial pancreas. Unlike other engineered tissues, the bioartificial pancreas does not need to physically integrate with the host's tissues, as its primary function is to release insulin in a controlled manner as a function of the patient's glucose levels. Furthermore, the typical patients needing insulin therapy do not have any functional islet tissue available; therefore, allogeneic or xenogeneic islet sources must be used. For these reasons, the bulk of studies on bioartificial pancreases use islets encapsulated in membrane-based devices protecting the islets from the recipient's immune system. The first devices (dating back to the 1970s) mostly consisted of hollow-fiber bioreactors in which islets were placed on the shell side of the bioreactor, and the patient's blood flowed through the fibers. Although these systems worked well over short periods of time, chronically implanted devices tended to activate blood clotting and eventually become clogged. More recently, more simple avenues have been explored, such as encapsulating islets in spherical or flat membranes placed in the host's tissues.

In theory, allogeneic implants will not trigger immune responses as long as the pore size of the capsule is small enough to prevent immune cells from the host from accessing the antigens expressed on the implanted cells. However, antigens shed from the cell surface may trigger a humoral immune response leading to the formation of antibodies. The proteins involved in this type of immune rejection include immunoglobins such as IgG and IgM, and complement molecules, the largest of which is C1q (mol wt = 410 kD, diameter = 30 nm) which is required for classical pathway complement-mediated cell damage. Binding of immunoglobins to cells in the implant without the presence of the complement system would not necessarily lead to cell damage; therefore, some investigators have claimed that a membrane with an effective pore size of less than 30 nm may suffice to protect the encapsulated cells. Smaller membrane pore sizes (<50 kD) have been typically used, however. This strategy has shown promise in animal models, with function remaining after several

months to a year, although the reliability of the procedure is far from perfect. A common recurrent problem with immuno-isolated cells is the presence of a foreign-body reaction against the capsule material itself, leading to the generation, over a period of days to weeks, of a fibrotic layer around it, compromising nutrient transport and the release of insulin from the implanted cells. Besides improvements in the biocompatibility of the material, one avenue that may improve function of these devices is the use of materials and/or factors promoting the growth of blood vessels near the surface of the capsule.

It has also been suggested that the longevity of islet cells may be limited in encapsulated systems, and that integration into the host tissue may be necessary for a permanent cure. Thus, as an alternative to immuno-isolation, other approaches are currently being sought to either eliminate the antigenic proteins and polysaccharide moieties on implanted cells or interfere with the signaling pathways governing these immune responses. These studies are in fact not limited to tissue-engineered constructs, but are also under investigation for the transplantation of whole organs. For example, transgenic strains of pigs, which have a body size similar to a human and which express human surface antigens, are currently being developed.

IV. FUTURE PROSPECTS FOR TISSUE ENGINEERING

Tissue engineering is a relatively new and rapidly evolving field still in its infancy. Exciting new discoveries in biology will soon open new avenues for tissue engineers. One of these discoveries is the recent identification of stem cells. Stem cells have a high replication potential and can differentiate into a large number of different cell types. The best characterized stem cells are those of hematopoietic origin that populate the bone marrow. These cells are also found in very small numbers in the peripheral circulation. They have been cultured successfully *in vitro* on a stromal layer of connective tissue cells to produce all common blood cell lineages, including red blood cells, monocytes, lymphocytes, and platelets. More recent discoveries suggest that wound healing in specialized tissues such as muscle sometimes involves the homing of stem cells present in the circulation. While the exact nature of these cells remains to be elucidated, they open up exciting avenues for tissue engineering. For example, such stem cells could be harvested from a patient's blood, (requiring a minimally invasive procedure), grown, differentiated *in vitro* into the tissue type needed, and then implanted back into the patient. Since patients would receive their own cells, no immune suppression would be needed.

Another source of stem cells is embryos, which can be obtained at the blastocyst stage. Embryonic stem cells are totipotent, meaning that they have the potential to differentiate into any cell type found in the body. In the presence of leukemia inhibitory factor, embryonic stem cells self-renew without any loss of development potential. Otherwise, they differentiate into a wide variety of cell types, the nature of which depends on the specific factors added to the culture medium. Cloning techniques enable replacing the original DNA from the embryonic stem cell with that of a patient (extracted from one of the patient's cells such as skin). The availability of such cells could have important implications for engineering tissues made of cells that have typically lost their ability to replicate, such as neurons, lung epithelium, etc. However, serious ethical considerations will have to be resolved prior to using human embryonic stem cells in such applications. Furthermore, more progress is needed in order to increase the yield of specific cell types used in tissue engineering from stem cells.

Clinical applications for engineered tissues often require a readily available supply of a large number of cells when the need arises. Maintaining a continuous supply by culture techniques or obtaining fresh cells in large numbers from animal or human sources is clearly impractical. Thus, long-term preservation methods will be critical for the future clinical applications of tissue engineering. Cryopreservation is the most efficient method of preservation, and careful studies of the effects of freezing-associated osmotic, chemical, thermal, and mechnical stresses will be required. Although many such studies have been carried out on dissociated cells in suspension, there have been few studies on tissue constructs, which pose special challenges because the optimal freezing conditions for different cell types may not be the same, and the freezing conditions may be difficult to control uniformly in a three-dimensional system.

In summary, tissue engineering encompasses a wide spectrum of disciplines, including biological and chemical sciences, engineering sciences, and medicine. Although tissue engineering is a relatively new field, exciting applications, varying from artificial skin to treat severe burns patients to a bioartificial pancreas to treat diabetics, have in some cases reached standard clinical practice, and in others shown major advances and promising preliminary clinical results. Thus, it is not unreasonable to expect that a number of new tissue engineering approaches will enter the realm of clinical applications within the next decade. However, it should be borne in mind that clinical success relies heavily on our fundamental understanding of the many complex issues associated with reconstruction and modification of tissues as well as the development of reliable technologies for large-scale handling of tissues.

SEE ALSO THE FOLLOWING ARTICLES

Biomaterials, Synthesis, Fabrication, and Applications • Bioreactors • Hybridomas, Genetic Engineering of • Mammalian Cell Culture • Metabolic Engineering • Polymers, Synthesis • Surface Chemistry

BIBLIOGRAPHY

Freshney, R. I. (2000). "Culture of Animal Cells. A Manual of Basic Technique," 4th ed., Wiley-Liss, New York.

Galletti, P. M., and Nerem, R. M., eds. (2000). "Prostheses and artificial organs," *In* The Biomedical Engineering Handbook, 2nd ed., Vol. 2, pp. 126.1–138.15, CRC Press, Boca Raton, FL.

Greco, R. S., ed. (1994). "Implantation Biology: The Host Response and Biomedical Devices," CRC Press, Boca Raton, FL.

Kreis, T., and Vale, R., eds. (1999). "Guidebook to the Extracellular Matrix, Anchor, and Adhesion Proteins," 2nd ed., Oxford University Press, London.

Lanza, R. P., Langer, R. S., and Vacanti, J. P., eds. (2000). "Principles of Tissue Engineering," 2nd ed., Academic Press, San Diego, CA.

Lauffenburger, D. A., and Lindermann, J. J. (1993). "Receptors: Models for Binding, Trafficking, and Signaling," Oxford University Press, London.

Lee, K., Berthiaume, F., Stephanopoulos, G. N., and Yarmush, M. L. (1999). "Metabolic flux analysis: a powerful tool for monitoring tissue function," *Tissue Eng.* **5,** 347–368.

Morgan, J. R., and Yarmush, M. L., eds. (1999). "Tissue Engineering Methods and Protocols," Humana Press, Totowa, NJ.

Palsson, B. Ø., and Hubbell, J. A., eds. (2000). "Tissue engineering," *In* "The Biomedical Engineering Handbook," 2nd ed., Vol. 2, pp. 109.1–125.17, CRC Press, Boca Raton, FL.

Patrick, C. W., Jr., Mikos, A. G., and McIntire, L. V., eds. (1998). "Frontiers in Tissue Engineering," Elsevier Science, New York.

Ratner, B. D., Hoffman, A. S., and Schoen, F., eds. (1997). "Biomaterials Science: An Introduction to Materials in Medicine," Academic Press, San Diego, CA.

Toxicology in Forensic Science

Olaf H. Drummer
Monash University

I. Applications of Forensic Toxicology
II. Specimens
III. Chain of Custody
IV. What Chemicals Should Be Targeted?
V. Techniques Used
VI. Initial Tests and Confirmation
VII. Quality Assurance and Validation
VIII. Reports
IX. Interpretation of Toxicological Results
X. Artefacts in Analysis
XI. Court Testimony and Expertise

GLOSSARY

Amphetamines A class of drugs that act as powerful stimulants with actions similar to norepinephrine.

Benzodiazepines A class of over 50 drugs that act as minor tranquilizers and hypnotics and include diazepam, temazepam, flunitrazepam, oxazepam, and alprazolam etc.

Central nervous system (CNS) The part of the body including the brain and spinal cord.

Confirmatory testing The use of a second test to confirm the intial test such that there is no doubt over the presence of the substance in the sample.

Ethical drugs Drugs that are available by prescription or legally over the counter.

Exhibits Items brought to a laboratory that form part of the evidence in a case; can be tissue specimens or physical items such as drug powders, tablets, syringes, etc.

Forensic toxicology The application of toxicology to the needs of the law.

Initial testing The use of screening tests to establish the likely presence of drugs or drug classes in a sample.

Opioids A class of drugs related to morphine, often also known as opiates.

TOXICOLOGY in the context of forensic science deals with the detection of drugs and other chemicals in situations involving legal proceedings. This discipline is best called *forensic toxicology*; however, the term *analytical toxicology* is also often applied to this subspecialty of toxicology. A forensic toxicologist is concerned with the detection of drugs or poisons in samples and is capable of

TABLE I Types of Specimens Collected and Their Principal Applications

Specimen	Application
Blood (or plasma or serum)	Most commonly used for drug detection and is the preferred single specimen for perpetrators and victims of crimes, death investigation cases, drivers of motor vehicles suspected of drug use
Urine	Used in all cases as a screening specimen for drugs of abuse, and is the preferred specimen for workplace, rehabilitation, and corrections drug testing programs, as well as in sports drug testing programs
Liver and other tissues from deceased persons	Used to supplement blood toxicology in some death investigations, particularly when body is decomposed and when the interpretation of blood results are equivocal
Hair	Used when a longer period of drug exposure is required (1–6 months) to complement other information on drug use
Sweat and saliva	Occasionally used as alternatives to other forms of testing when onsite[a] results are required in workplace and correctional or rehabilitation settings
Breath	Used extensively to establish presence of alcohol in drivers of motor vehicles and in workplace settings

[a] Onsite refers to an initial detection of drugs in specimens at the point of collection, rather than waiting for a result from a laboratory.

defending the results in a court of law. This distinction from an ordinary analytical toxicologist is important, as a conventional toxicologist is mainly concerned with the detection of substances and may not understand the specific medico-legal requirements in forensic cases. A forensic toxicologist also is able to assist legal proceedings in the interpretation of the significance of the results obtained.

I. APPLICATIONS OF FORENSIC TOXICOLOGY

Forensic toxicology has a number of applications. It provides clinicians with information of a possible drug taken in overdose or authorities investigating a sudden death or poisoning with the possible substances(s) used.

Toxicology testing is also important in victims of crime, or in persons apprehended for a crime. Drugs may have been given by the assailant to reduce consciousness of the victim, such as in rape cases. These drugs include the benzodiazepines (e.g., Rohypnol, Valium, Ativan, etc.) and gamma-hydroxybutyrate (GHB). Toxicology also establishes if any drug was used by the victim that may have affected consciousness or behavior. Defendants arrested shortly after allegedly committing a violent crime may be under the influence of drugs. Alcohol and drugs are also commonly targeted in drivers suspected of driving under the influence.

Ultimately, toxicology testing results will assist the investigator, pathologist, coroner, or medical examiner in establishing evidence of drug use or refuting the use of relevant drugs. This latter application is essential, as few drugs leave any visible trace of their presence in a person.

Forensic toxicology is also used in employment drug testing, rehabilitation settings, and in human performance testing. The detection of drugs of abuse in potential employees prior to being hired is becoming an important application of toxicology. Attempts to exclude drugs from prisons and to aid in rehabilitation of drug-dependent persons are other applications of toxicology. Human performance testing relates to the detection of drugs that might have improved performance in athletic events. This testing may even apply to racing animals such as horses, camels, dogs, etc. Specimens used in these cases are usually urine, although hair is being increasingly used to provide a greater window of opportunity in workplace settings.

II. SPECIMENS

A wide variety of specimens can be collected to assess possible drug use. These include blood (or sometimes plasma or serum*), urine, breath, saliva, sweat in living persons, or a range of other tissues from bodies at autopsy during death investigations. In cadavers the most common specimens, after blood and urine, are liver, bile, and vitreous humour. Muscle, brain, bone, and fat do have uses in certain types of cases when blood is unavailable due to decomposition or following suspicion of unusual poisons. The principal applications of some specimens are summarized in Table I.

III. CHAIN OF CUSTODY

Courts and other legal processes usually require proof that the laboratory has taken all reasonable precautions against unwanted tampering or alteration of the evidence. This applies to specimens and to physical exhibits used by the laboratory for toxicology investigations. Consequently, it is essential that the correct identifying details are recorded on the exhibit or specimen container and an adequate record is kept of persons in possession of the exhibit(s). Alternatively, when couriers are used to transport exhibits, the exhibit must be adequately sealed to prevent unauthorized tampering.

*Obtained from blood by either centrifugation to remove red blood cells (plasma) or other types of separation processes to obtain plasma or serum (fluid obtained after blood has clotted).

TABLE II Drugs Commonly Targeted in Forensic Toxicology Investigations

Most common drugs	Alcohol
	Amphetamines, benzodiazepines, cannabis, cocaine and opiates
Common ethical drugs	Antidepressants including tricyclics, serotonin reuptake inhibitors, and monoamine oxidase inhibitors
	Antipsychotics drugs such as phenothiazines, haloperidol, olanzapine, or clozapine
	Other analgesics and anti-inflammatory agents
	Digoxin, anti-arrhythmic drugs, anti-hypertensive drugs, and many other cardiovascular drugs
Less common drugs	GHB, LSD, and other hallucinogens
	Anabolic steroids and other performance-enhancing drugs
	Barbiturates and older sedatives and hypnotics such as methaqualone
	Volatile substances such as butane, or gasoline
	Various domestic, industrial, and agricultural poisons such as organophosphates, or solvents

IV. WHAT CHEMICALS SHOULD BE TARGETED?

It is common to find a variety of ethical and illicit drugs or unusual poisons. Worldwide experience also shows that forensic cases often involve more than one drug substance. High rates of multiple drug use are found in deaths from misuse of drugs and also in perpetrators of violent crimes.

It is also well known by forensic toxicologists that the information provided to the laboratory concerning possible drug use may not agree with what is actually detected. It is therefore strongly recommended that laboratories provide a systematic approach to their analyses and include as wide a range of common ethical and illicit drugs as feasible. A laboratory using this approach would normally include a range of screening methods often incorporating both chromatographic and immunological techniques. Drug classes such as alcohol, analgesics, opioid and non-opioid narcotics, amphetamines, antidepressants, benzodiazepines, barbiturates, cannabis, cocaine, major tranquilizers (antipsychotic drugs), and other CNS-depressant drugs should be included (Table II).

The incorporation of a reasonably complete range of drugs in any testing protocol is important, as many of these drugs are mood altering and can therefore affect behavior as well as the health of an individual. Persons using benzodiazepines, for example, will be further affected by alcohol and other CNS-active drugs. The toxic concentrations of drugs are also influenced by the presence of other potentially toxic drugs. For example, the toxicity of heroin is affected by the concomitant use of alcohol and other CNS-depressant drugs.

V. TECHNIQUES USED

The range of techniques available to detect drugs in specimens or physical exhibits varies from commercial kit-based immunoassays and traditional thin-layer chromatography (TLC) to instrumental separation techniques such as high-performance liquid chromatography (HPLC), gas chromatography (GC), and capillary electrophoresis (CE). Mass spectrometry (MS) is the definitive technique used to establish proof of structure of an unknown substance and can be linked to GC, HPLC, and more recently to CE.

The use of appropriate extraction techniques is critical to all analytical methods. Three main types of extractions are used: liquid–liquid, solid-phase, and direct injection. Traditionally, liquid techniques have been favored in which a blood or urine specimen is treated with a buffer of an appropriate pH followed by a solvent capable of partitioning the drug out of the matrix. Solvents used include chloroform, diethyl ether, ethyl acetate, toluene, hexane, various alcohols, and butyl chloride and mixtures thereof. The solvent is then isolated from the mixture and either cleaned up by another extraction process or evaporated to dryness.

Solid-phase techniques are becoming increasingly favored, as they offer the ability to extract substances of widely differing polarity more readily than with liquid techniques.

Direct-injection techniques into either GC or HPLC instruments bypass the extraction step and can offer a very rapid analytical process. In GC, solid-phase microextraction (SPE) can be used, while HPLC tends to require use of pre-columns that are backflushed with the use of column-switching valves.

VI. INITIAL TESTS AND CONFIRMATION

The process of conducting toxicology in forensic science is similar to other analytical disciplines, in that sufficiently suitable analytical techniques need to be employed that are appropriately validated. The foremost goal is the need to provide a substantial proof of the presence of a substance(s). The use of conventional GC, TLC, or HPLC by themselves would not normally be sufficient to accept unequivocal proof of the presence of a chemical substance.

Two or more independent tests are normally required, or the use of a more powerful analytical test, such as mass spectrometry (MS) may often be preferred. Because of the need to perform a rigorous analysis, the analytical schema is often broken up into two steps. The identification stage is termed the screening or initial test, while the second analytical test is the confirmation process. The confirmation process often also provides a quantitative measure of how much substance was present in the sample; otherwise, a separate test is required to quantify the amount of substance present in the specimen (see later).

In all processes, it is important that no analytical inconsistency appears, else a result may be invalidated. For example, in the identification of amphetamine in a blood specimen, an immunoassay positive to the amphetamine class is expected to be positive for one or more amphetamines in the confirmation assay. The apparent detection of a drug in one analytical assay but not in another means that the drug was not confirmed, providing both assays are capable of detecting this drug or one or more members of a drug class.

While MS is the preferred technique for confirmation of drugs and poisons, some substances display poor mass spectral definition. Compounds with base ions at mass/charge ratios of less than 100 or with common ions such as m/z 105 and with little or no ions in the higher mass range are not recommended for confirmation by MS alone. Derivitization of a functional group to produce improved mass spectral properties can often be successful. Common derivatives include acyl esters, silyl ethers, etc. Alternatively, reliance on other chromatographic procedures can provide adequate confirmation. It is important when using any chromatographic procedure (HPLC, GS, CE, etc.) that the retention time of the substance being identified matches that of an authentic standard.

Some apparent analytical inconsistencies may provide important forensic information. For example, if a result for opiates is negative in urine, but positive in blood, it is quite likely that heroin* was administered shortly before death and the metabolites had not yet been excreted. This situation is often found in acute sudden death among heroin users in whom substantial urinary excretion has not yet occurred.

VII. QUALITY ASSURANCE AND VALIDATION

Essential components of any form of toxicological testing are validation and quality assurance. It is important that the testing method used is appropriately validated; that is, it has been shown to accurately and precisely detect particular substance(s), there is little or interference from other drugs or from the matrix, and a useful detection limit has been established. Moreover, it is essential that the method is rugged and will allow any suitably trained analyst to conduct the procedure and achieve the same results as another analyst. To achieve these aims, it will be necessary to trial the method in the laboratory with specimens of varying quality before full validation can be achieved.

It is recommended to include internal quality controls with each batch of samples to enable an internal check of the reliability of each assay. These controls contain known drugs at known concentrations. Suitable acceptance criteria are required for these controls before results of unknown cases can be accepted and released to a client. Acceptance criteria vary depending on the analyte and application. For example, blood alcohol estimations have acceptance criteria less than 5%,* while postmortem blood procedures may be 10 to 20%.

An important feature of analytical assays in forensic toxicology is the use of internal standards. These are drugs with chemical and physical characteristics similar to the drug(s) being analyzed and, when added at the start of the extraction procedure, provide an ability to negate the effects of variable or low recoveries from the matrix. Hence, even when recoveries are low, the ratios of analyte and drug are essentially the same as for situations of higher recovery. An ideal recovery marker is when the internal standard is a deuterated analog of the analyte. When deuterated internal standards are used, it may not be necessary to match the calibration standards with the same matrix as the unknown samples. It is important, however, that absolute recoveries are reasonable (i.e., at least over 30%). This ensures less variability between samples and optimizes the detection limit.

From time to time, it will be important to run unknown samples prepared by another laboratory or by a person not directly involved in laboratory work to establish proficiency. These are known as proficiency programs or quality assurance programs. These trials are often conducted with many other laboratories conducting similar work and provide an independent assessment of the proficiency of the laboratory to detect (and quantify) specific drugs. The performance of the laboratory should be regularly assessed from these results and any corrective action implemented, if appropriate. This process provides a measure of continuous improvement, an essential characteristic of any laboratory.

*Heroin is rapidly metabolized to morphine.

*Normally, the coefficient of variation (CV) of the mean is calculated as a standard deviation divided by the mean of the result.

VIII. REPORTS

Once an analysis is complete a report must be issued to the client(s) which accurately details the analytical findings. These results should indicate the type of tests conducted, the analytical method used (HPLC, GC–MS, etc.), on which specimens the analyses were conducted, and of course the result(s). The result(s) should be unambiguous, using such terms as "detected" or "not detected." The use of the term "not present" should be avoided, as it implies no possibility of the substance being present. A toxicologist can rarely be so definitive and can only indicate that a substance was not detected at a certain threshold concentration. For this reason, a detection limit alongside tests for specific substances should be provided for "not detected" results.

For quantitative results consistency in units is advised and should not be given with more significant digits than the accuracy will allow. For example, there is no point in reporting a result for blood morphine as 0.162 mg/L when the accuracy and precision of the method is ±20%. A result of 0.16 mg/L would suffice.

For drug-screening results, it is advisable to provide clients with an indication of the range of substances a method is capable of detecting and some indication of the detection limit, such as "at least therapeutic concentrations" or "only supra-therapeutic concentrations."

In postmortem cases, all reports should indicate the site of blood sampling and provide (where relevant) some comment on the possibility of postmortem artefacts such as redistribution. By incorporating these comments, those reading the report are less likely to unwittingly misinterpret the results.

IX. INTERPRETATION OF TOXICOLOGICAL RESULTS

Interpretation of any toxicological result is complex. Consideration must be given to the circumstances of the case, and in particular what significance may be drawn from the toxicology. For example, the finding of a drug in potentially toxic concentrations in a person killed by a gunshot wound to the head cannot reasonably lead to the conclusion that the drug caused the death. On the other hand, the absence of an obvious anatomical cause of death will lead investigators to consider the role of any drug use. Considerations must include the chronicity of drug use, the likely time of ingestion, the route of ingestion, the age and health of the person (e.g., presence of heart, liver, or kidney disease), the use of other active substances, and even genetic factors that may lead to an altered metabolism.

X. ARTEFACTS IN ANALYSIS

A. Stability of Drugs

Chemical instability occurs for a number of drugs and metabolites that will alter the concentration and even cause the drug to disappear if storage conditions are not optimal. This will occur at room temperature and even sometimes when specimens are stored frozen at −20°C.

Alcohol will be lost to evaporation unless sealed tubes are used or specimens are stored at −80°C; however, alcohol can also be produced by bacterial action on glucose and other sugars found in blood. The use of potassium fluoride as a preservative (minimum 1% w/v) is required to prevent bacterial activity for up to one month after collection when the sample is stored at 4°C.

B. Bioconversion

A number of drugs can undergo chemical changes in a body after death. These chemical changes can be either metabolically mediated or caused by spontaneous degradative processes. For example, the metabolism of heroin to morphine occurs in life, in blood and other tissues following collection, or even *in situ* when a person has died. For this reason, heroin or the immediate 6-acetyl morphine are rarely detected in blood. Morphine is therefore the target drug. Aspirin is also converted rapidly to salicylate by hydrolytic mechanisms. Most drugs activated by de-esterification or hydrolysis will be subject to similar processes.

Nitro-containing drugs, such as the benzodiazepines, nitrazepam, or flunitrazepam, are also rapidly biotransformed after death to their respective amino metabolites by the action of certain types of bacteria. Toxicologists must therefore target their analyses to these transformation products rather than the parent drug.

Sulfur-containing drugs, such as dothiepin, thiopental, or thioridazine, are also subject to bacterial attack during the postmortem interval, leading to progressive losses over time. Of course, the parallel process of tissue loss will also affect the tissue concentration during putrefaction.

C. Redistribution

The process of death imparts a number of other special processes that affect the collection and analysis of specimens obtained at autopsy. These include postmortem redistribution in which the concentration of a drug in blood has been affected by diffusion of drug from neighboring tissue sites and organs such as stomach contents. This is minimized, but not arrested, by using peripheral blood from the femoral region. Even liver concentrations are affected

TABLE III Likely Extent of Postmortem Redistribution for Selected Drugs[a]

Drug/drug class	Likely extent of postmortem redistribution
Acetaminophen (paracetamol)	Low
Alcohol (ethanol)	Low
Barbiturates	Low to moderate
Benzodiazepines	Low to moderate
Cocaine	Low
Digoxin	Very high
Methadone	Low to moderate
Morphine	Low
Phenothiazines	Moderate to high
Propoxyphene	Very high
Salicylate	Low
Serotonin reuptake inhibitors	Low to moderate
Tricyclic antidepressants	High

[a] These changes should only be used as a guide, as the environmental conditions, length of time from death to specimen collection, and quality of specimen can affect the extent of redistribution.

Note: Low = up to 20% elevation; moderate = 21–50%; high = 50–200%; very high > 200%.

by diffusion from intestinal contents or from incomplete circulation and distribution within the liver.

This process is particularly significant for drugs with high lipid solubility, as these drugs tend to show concentration differences in tissues and blood. Table III shows the extent of these changes for selected drugs when comparisons are made between blood collected from the heart and that collected from the femoral region.

The femoral blood is least subject to redistribution after death; however, drugs with much higher concentrations in muscular tissue will still diffuse through the vessel walls and elevate the neighboring blood concentrations. If the femoral vessels are not tied off from the vena cava and aorta, then the process of drawing blood can also extract blood from the abdominal cavity that has been contaminated from diffusion of gastric and intestinal contents. It is therefore advisable to reduce these processes by collecting blood specimens as soon as possible after death from the femoral region with blood vessels tied off to reduce contamination.

XI. COURT TESTIMONY AND EXPERTISE

Forensic toxicologists and other professionals called to give evidence in court should consider that much of their technical evidence is beyond the ready comprehension of lay people in juries, legal counsel, and judges. Restricting one's testimony to understandable language and simple concepts is highly recommended.

A further problem relates to an assumption often made by legal counsel (and indeed other parties) that a toxicological investigation was exhaustive and all drugs and poisons were excluded in the testing processes. Most toxicology performed is restricted to a few analytical tests for a range of "common drugs and poisons," unless the client has made a request to examine for (additional) specific chemicals. Analysts should make courts aware of the actual testing conducted and provide a list of substances incorporated in the investigation. Importantly, advice on any limitations applied to the interpretation of the analytical results should be provided (e.g., poor-quality specimens or postmortem artifacts). Above all, toxicologists must restrict their evidence to those areas for which they claim expertise. Stretching their expertise to apparently assist the court can lead to incorrect or misleading evidence and damage the reputation of the expert.

SEE ALSO THE FOLLOWING ARTICLES

Analytical Chemistry • DNA Testing in Forensic Science • Environmental Toxicology • Mass Spectrometry in Forensic Science • Organic Chemistry, Compound Detection • Spectroscopy in Forensic Science

BIBLIOGRAPHY

Baselt, R. H., and Cravey, R. H. (1996). "Disposition of Toxic Drugs and Chemicals in Man," 4th ed., Year Book Medical Publishers, Chicago.

de Zeeuw, R. A. (1997). "Drug screening in biological fluids: the need for a systematic approach," *J. Chromatogr.* **689,** 71–79.

Drummer, O. H. (1998). "Adverse drug reactions." *In* "The Inquest Handbook" (H. Selby, ed.), The Federation Press, Leichhardt, NSW Australia.

Drummer, O. H. (1999). "Review: chromatographic screening techniques in systematic toxicological analysis," *J. Chromatogr.* **733,** 27–45.

Freckleton, I., and Selby, H. (1993). "Expert Evidence," LBS Information Services, Sydney, Australia.

International Association of Forensic Toxicologists (TIAFT). (2001). http://www.tiaft.org.

Karch, S. (1998). "Drug Abuse Handbook," CRC Press, Boca Raton, FL.

Levine, B. (1999). "Principles of Forensic Toxicology," AACC Press, Washington, D.C.

Maurer, H. H. (1992). "Systematic toxicological analysis of drugs and their metabolites by gas chromatography-mass spectrometry," *J. Chromatogr.* **118,** 3–42.

Moffatt, A. C., ed. (1986). "Clarke's Isolation and Identification of Drugs," The Pharmaceutical Press, London.

Siegel, J., ed. (2000). "Encyclopedia of Forensic Science," Academic Press, London.

Society of Forensic Toxicologists (SOFT). (2001). http://www.soft-tox.org.

United Nations. (1995). "Recommended Methods for the Detection

and Assay of Heroin, Cannabinoids, Cocaine, Amphetamine, Methamphetamine and Ring-Substituted Derivatives in Biological Specimens," U.N. Publ. No. ST/NAR/27, United Nations International Drug Control Programme, Vienna, Austria.

United Nations. (1997). "Recommended Methods for the Detection and Assay of Barbiturates and Benzodiazepines in Biological Specimens," U.N. Publ. No. ST/NAR/28, United Nations International Drug Control Programme, Vienna, Austria.